Appendix A: Supporting Figures

Summary

This section contains supporting figures for results presented in Chapter 3 and Chapter 4, as follows:

Appendix A.1: Correlation Plots for Models Described in Chapter 2

Appendix A.2: Parameter Estimation Analysis from Chapter 3 Section 3.1.3 performed with Different Threshold Values

Appendix A.3: Correlation Plots for Models Described in Chapter 4

Appendix A.4: Parameter Estimation Analysis from Chapter 4 Section 4.2.3 performed with Different Threshold Values

A.1 Correlation Plots for Models Described in Chapter 2

Figure A.1: Binned scatterplots (below diagonal) and correlation values (above diagonal; black: Spearman's rank correlation coefficient R^2, blue: associated p-value) of pairwise combinations of all summary statistics and model parameters using 120,000 successful simulations for the NM SW FLAT model.

The cells in each pairwise scatterplot (below diagonal) are shaded according to the number of data points they contain, with darker cells containing a larger number of data points than lighter cells. Pairwise panels giving the correlation values (above diagonal) are shaded according to R^2 value and the text greyed out for those that are not significant (p-value ≥ 0.05). P1: probability of cultural mutation; S2: probability of fission / extinction; P3: migration distance (km); P4: total number of items in each group's cultural repertoire; P5: maximum number of groups; S1: mean(SI:ornaments); S2: var(SI:ornaments); S3: mean(SI:sites); S4: var(SI:sites); S5: mean(MI:ornaments); S6: var(MI:ornaments); S7: mean(MI:sites); S8: var(MI:sites); S9: mean(DR); S10: var(DR); S11: MAD:ornaments; S12: MAD:sites.

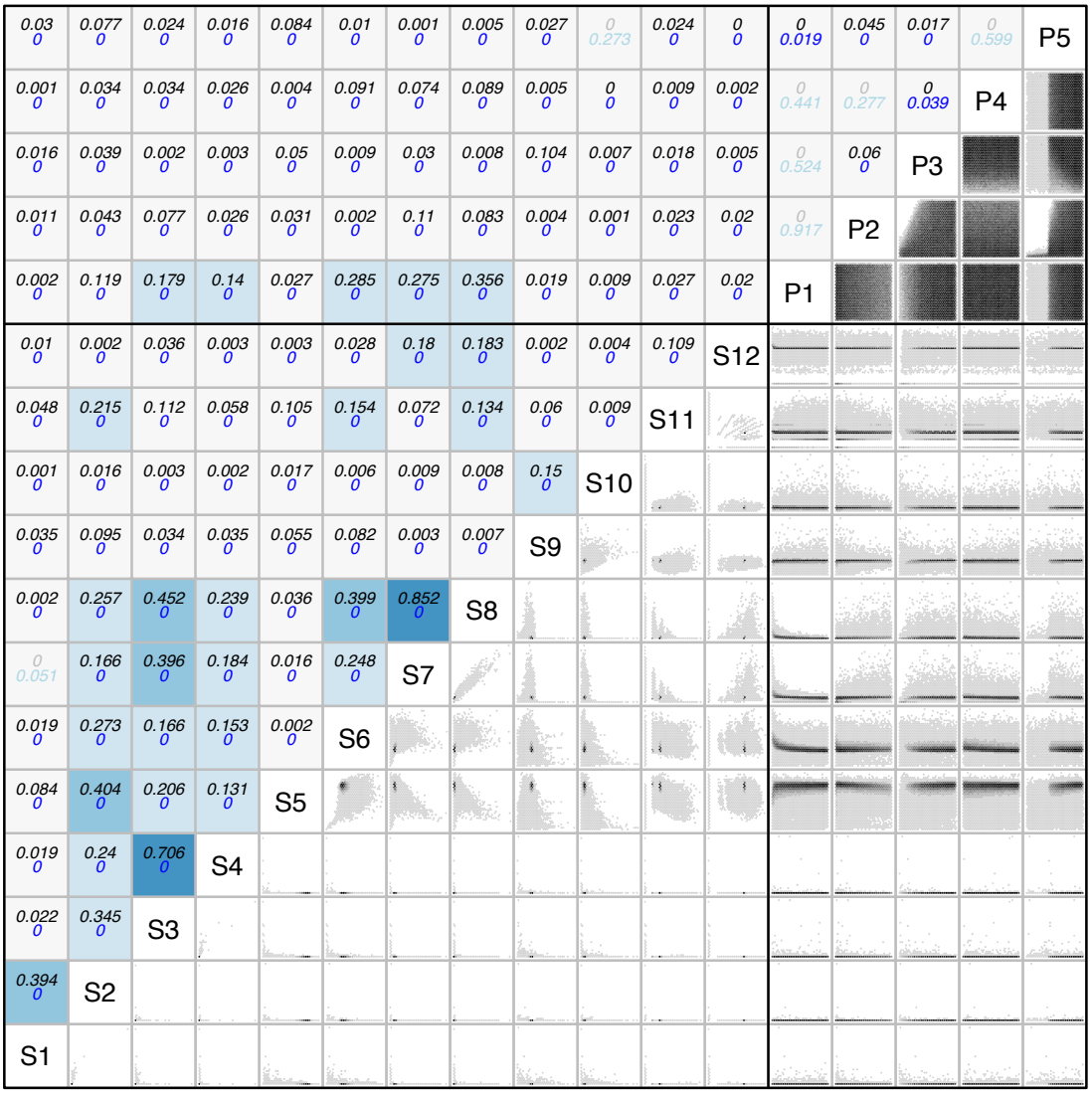


Figure A.2: Binned scatterplots (below diagonal) and correlation values (above diagonal; black: Spearman's rank correlation coefficient R^2, blue: associated p-value) of pairwise combinations of all summary statistics and model parameters using 120,000 successful simulations for the NM SW B-A model.

The cells in each pairwise scatterplot (below diagonal) are shaded according to the number of data points they contain, with darker cells containing a larger number of data points than lighter cells. Pairwise panels giving the correlation values (above diagonal) are shaded according to R^2 value and the text greyed out for those that are not significant (p-value ≥ 0.05). P1: probability of cultural mutation; S2: probability of fission / extinction; P3: migration distance (km); P4: total number of items in each group's cultural repertoire; P5: maximum number of groups; S1: mean(SI:ornaments); S2: var(SI:ornaments); S3: mean(SI:sites); S4: var(SI:sites); S5: mean(MI:ornaments); S6: var(MI:ornaments); S7: mean(MI:sites); S8: var(MI:sites); S9: mean(DR); S10: var(DR); S11: MAD:ornaments; S12: MAD:sites.

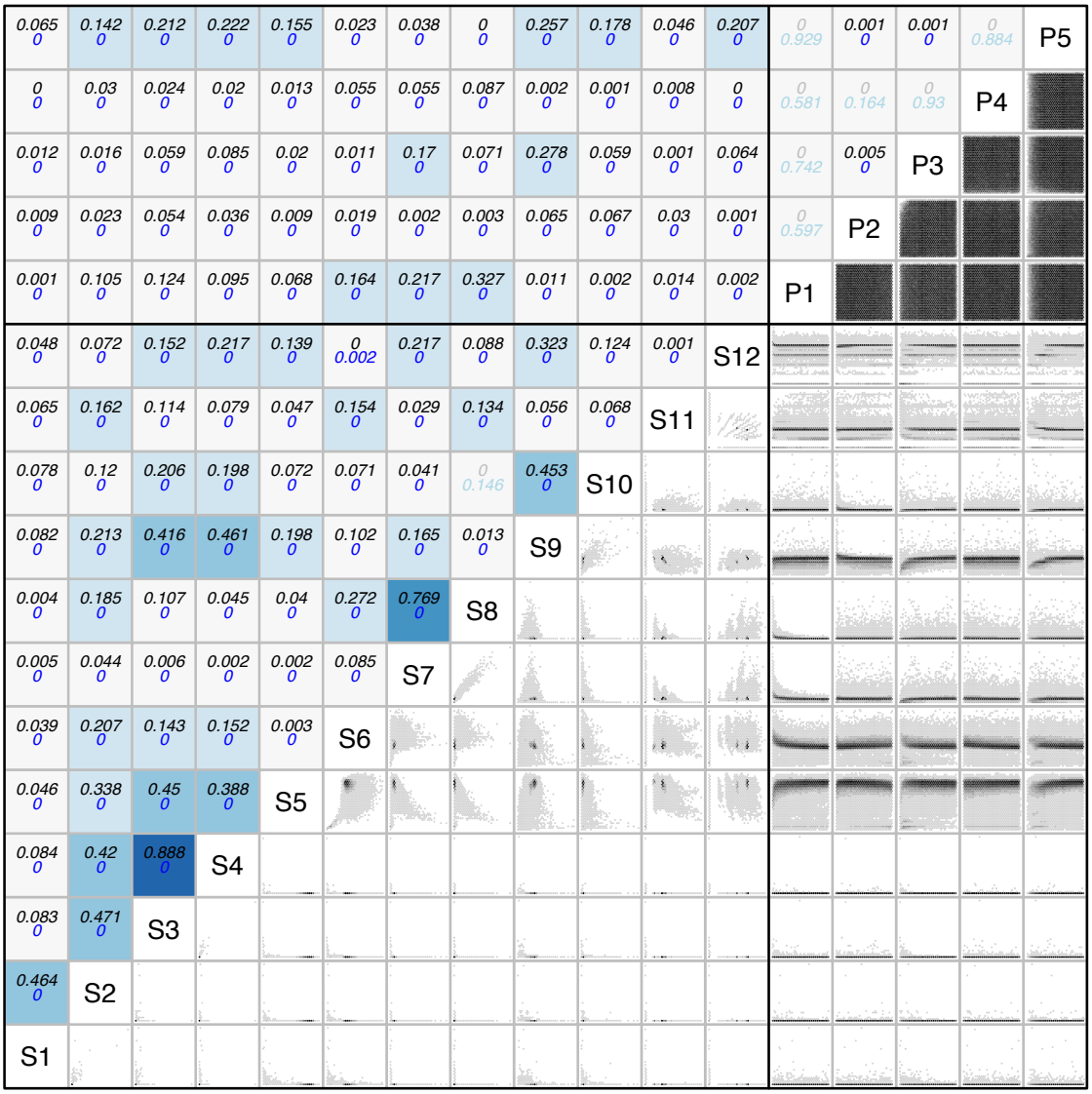


Figure A.3: Binned scatterplots (below diagonal) and correlation values (above diagonal; black: Spearman's rank correlation coefficient R^2, blue: associated p-value) of pairwise combinations of all summary statistics and model parameters using 400,000 successful simulations for the NMCI SW FLAT model.

The cells in each pairwise scatterplot (below diagonal) are shaded according to the number of data points they contain, with darker cells containing a larger number of data points than lighter cells. Pairwise panels giving the correlation values (above diagonal) are shaded according to R^2 value and the text greyed out for those that are not significant (p-value ≥ 0.05). P1: probability of cultural mutation; S2: probability of fission / extinction; P3: migration distance (km); P4: total number of items in each group's cultural repertoire; P5: maximum number of groups; P6: interaction radius (km); S1: mean(SI:ornaments); S2: var(SI:ornaments); S3: mean(SI:sites); S4: var(SI:sites); S5: mean(MI:ornaments); S6: var(MI:ornaments); S7: mean(MI:sites); S8: var(MI:sites); S9: mean(DR); S10: var(DR); S11: MAD:ornaments; S12: MAD:sites.

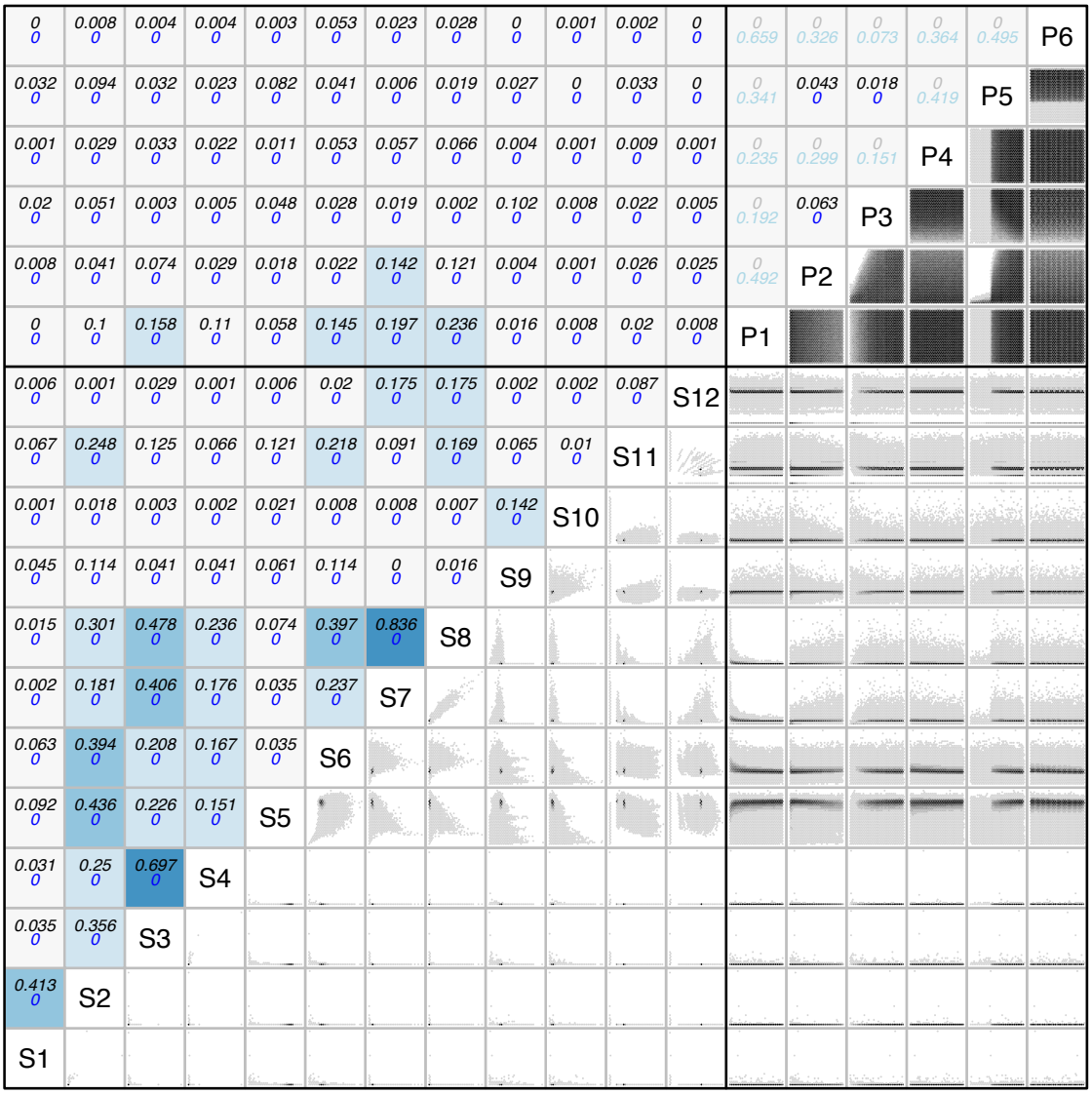


Figure A.4: Binned scatterplots (below diagonal) and correlation values (above diagonal; black: Spearman's rank correlation coefficient R^2, blue: associated p-value) of pairwise combinations of all summary statistics and model parameters using 400,000 successful simulations for the CD SW FLAT model.

The cells in each pairwise scatterplot (below diagonal) are shaded according to the number of data points they contain, with darker cells containing a larger number of data points than lighter cells. Pairwise panels giving the correlation values (above diagonal) are shaded according to R^2 value and the text greyed out for those that are not significant (p-value ≥ 0.05). P1: probability of cultural mutation; S2: probability of fission / extinction; P3: migration distance (km); P4: total number of items in each group's cultural repertoire; P5: maximum number of groups; P6: interaction radius (km); S1: mean(SI:ornaments); S2: var(SI:ornaments); S3: mean(SI:sites); S4: var(SI:sites); S5: mean(MI:ornaments); S6: var(MI:ornaments); S7: mean(MI:sites); S8: var(MI:sites); S9: mean(DR); S10: var(DR); S11: MAD:ornaments; S12: MAD:sites.

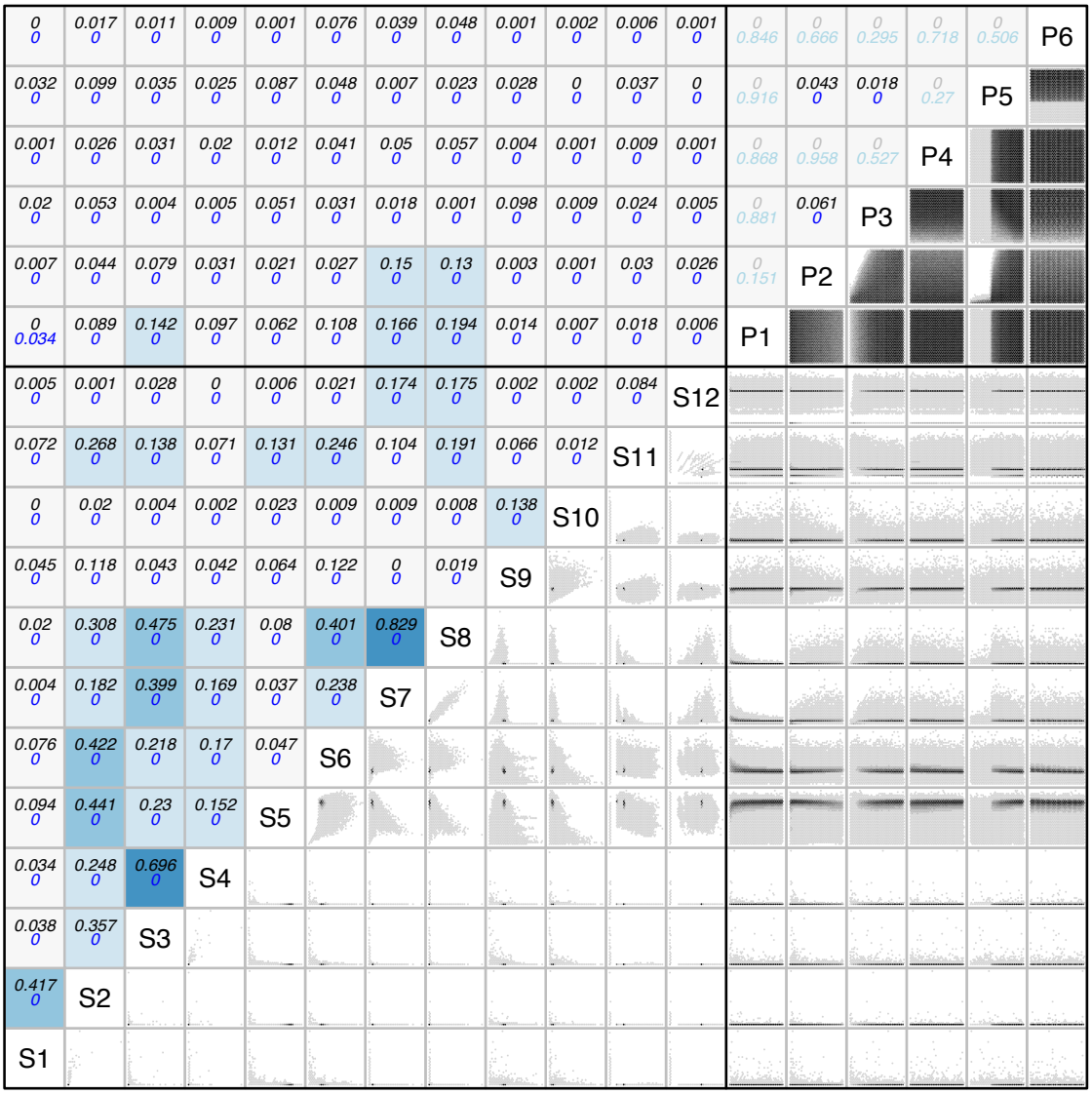


Figure A.5: Binned scatterplots (below diagonal) and correlation values (above diagonal; black: Spearman's rank correlation coefficient R^2, blue: associated p-value) of pairwise combinations of all summary statistics and model parameters using 400,000 successful simulations for the NMCI SW B-A model.

The cells in each pairwise scatterplot (below diagonal) are shaded according to the number of data points they contain, with darker cells containing a larger number of data points than lighter cells. Pairwise panels giving the correlation values (above diagonal) are shaded according to R^2 value and the text greyed out for those that are not significant (p-value ≥ 0.05). P1: probability of cultural mutation; S2: probability of fission / extinction; P3: migration distance (km); P4: total number of items in each group's cultural repertoire; P5: maximum number of groups; P6: interaction radius (km); S1: mean(SI:ornaments); S2: var(SI:ornaments); S3: mean(SI:sites); S4: var(SI:sites); S5: mean(MI:ornaments); S6: var(MI:ornaments); S7: mean(MI:sites); S8: var(MI:sites); S9: mean(DR); S10: var(DR); S11: MAD:ornaments; S12: MAD:sites.

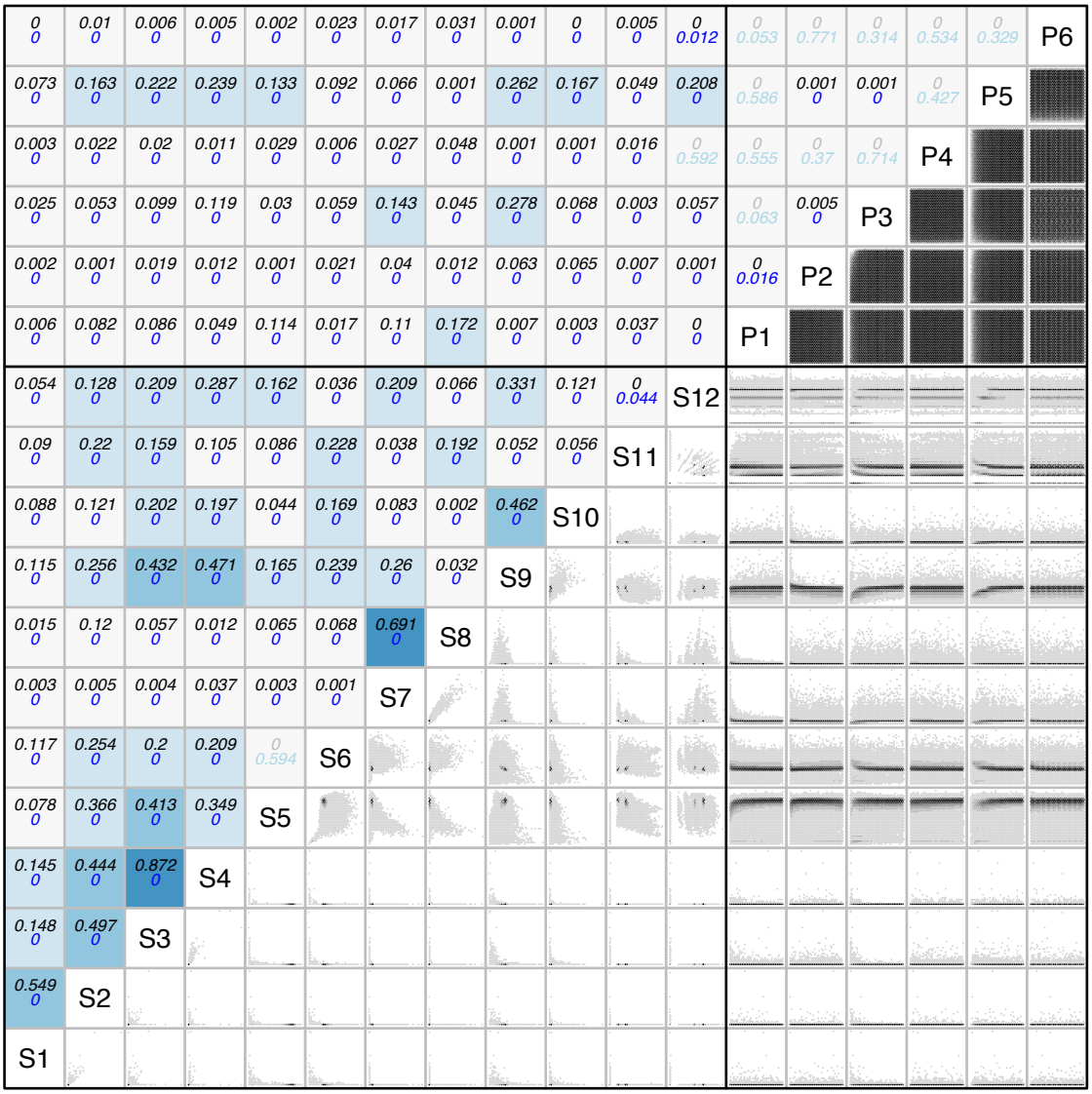


Figure A.6: Binned scatterplots (below diagonal) and correlation values (above diagonal; black: Spearman's rank correlation coefficient R^2, blue: associated p-value) of pairwise combinations of all summary statistics and model parameters using 400,000 successful simulations for the CD SW B-A model.

The cells in each pairwise scatterplot (below diagonal) are shaded according to the number of data points they contain, with darker cells containing a larger number of data points than lighter cells. Pairwise panels giving the correlation values (above diagonal) are shaded according to R^2 value and the text greyed out for those that are not significant (p-value ≥ 0.05). P1: probability of cultural mutation; S2: probability of fission / extinction; P3: migration distance (km); P4: total number of items in each group's cultural repertoire; P5: maximum number of groups; P6: interaction radius (km); S1: mean(SI:ornaments); S2: var(SI:ornaments); S3: mean(SI:sites); S4: var(SI:sites); S5: mean(MI:ornaments); S6: var(MI:ornaments); S7: mean(MI:sites); S8: var(MI:sites); S9: mean(DR); S10: var(DR); S11: MAD:ornaments; S12: MAD:sites.

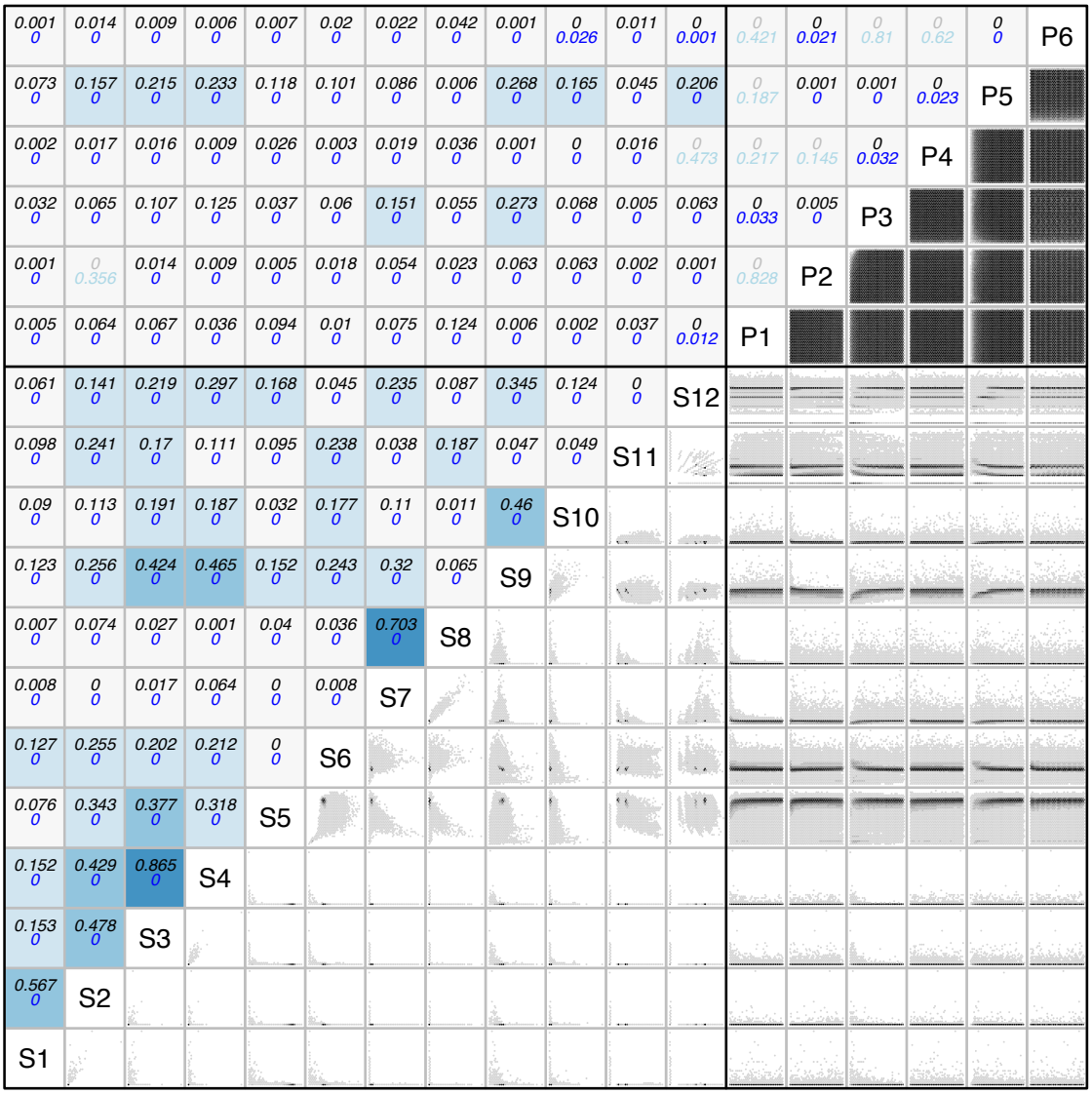


Figure A.7: Binned scatterplots (below diagonal) and correlation values (above diagonal; black: Spearman's rank correlation coefficient R^2, blue: associated p-value) of pairwise combinations of all summary statistics and model parameters using 400,000 successful simulations for the NMCI DIS FLAT model.

The cells in each pairwise scatterplot (below diagonal) are shaded according to the number of data points they contain, with darker cells containing a larger number of data points than lighter cells. Pairwise panels giving the correlation values (above diagonal) are shaded according to R^2 value and the text greyed out for those that are not significant (p-value ≥ 0.05). P1: probability of cultural mutation; S2: probability of fission / extinction; P3: migration distance (km); P4: total number of items in each group's cultural repertoire; P5: maximum number of groups; P6: interaction radius (km); S1: mean(SI:ornaments); S2: var(SI:ornaments); S3: mean(SI:sites); S4: var(SI:sites); S5: mean(MI:ornaments); S6: var(MI:ornaments); S7: mean(MI:sites); S8: var(MI:sites); S9: mean(DR); S10: var(DR); S11: MAD:ornaments; S12: MAD:sites.

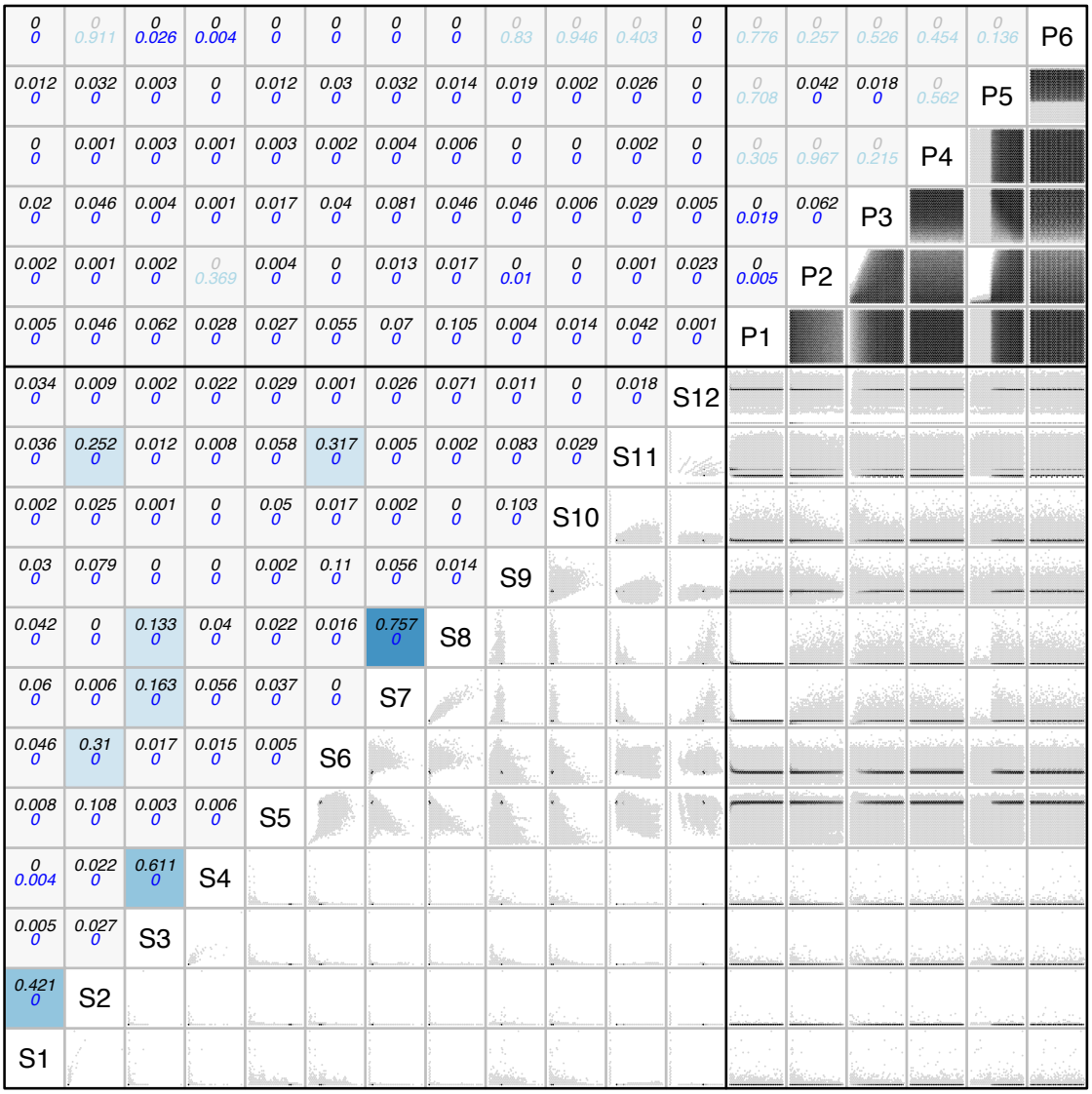


Figure A.8: Binned scatterplots (below diagonal) and correlation values (above diagonal; black: Spearman's rank correlation coefficient R^2, blue: associated p-value) of pairwise combinations of all summary statistics and model parameters using 400,000 successful simulations for the CD DIS FLAT model.

The cells in each pairwise scatterplot (below diagonal) are shaded according to the number of data points they contain, with darker cells containing a larger number of data points than lighter cells. Pairwise panels giving the correlation values (above diagonal) are shaded according to R^2 value and the text greyed out for those that are not significant (p-value ≥ 0.05). P1: probability of cultural mutation; S2: probability of fission / extinction; P3: migration distance (km); P4: total number of items in each group's cultural repertoire; P5: maximum number of groups; P6: interaction radius (km); S1: mean(SI:ornaments); S2: var(SI:ornaments); S3: mean(SI:sites); S4: var(SI:sites); S5: mean(MI:ornaments); S6: var(MI:ornaments); S7: mean(MI:sites); S8: var(MI:sites); S9: mean(DR); S10: var(DR); S11: MAD:ornaments; S12: MAD:sites.

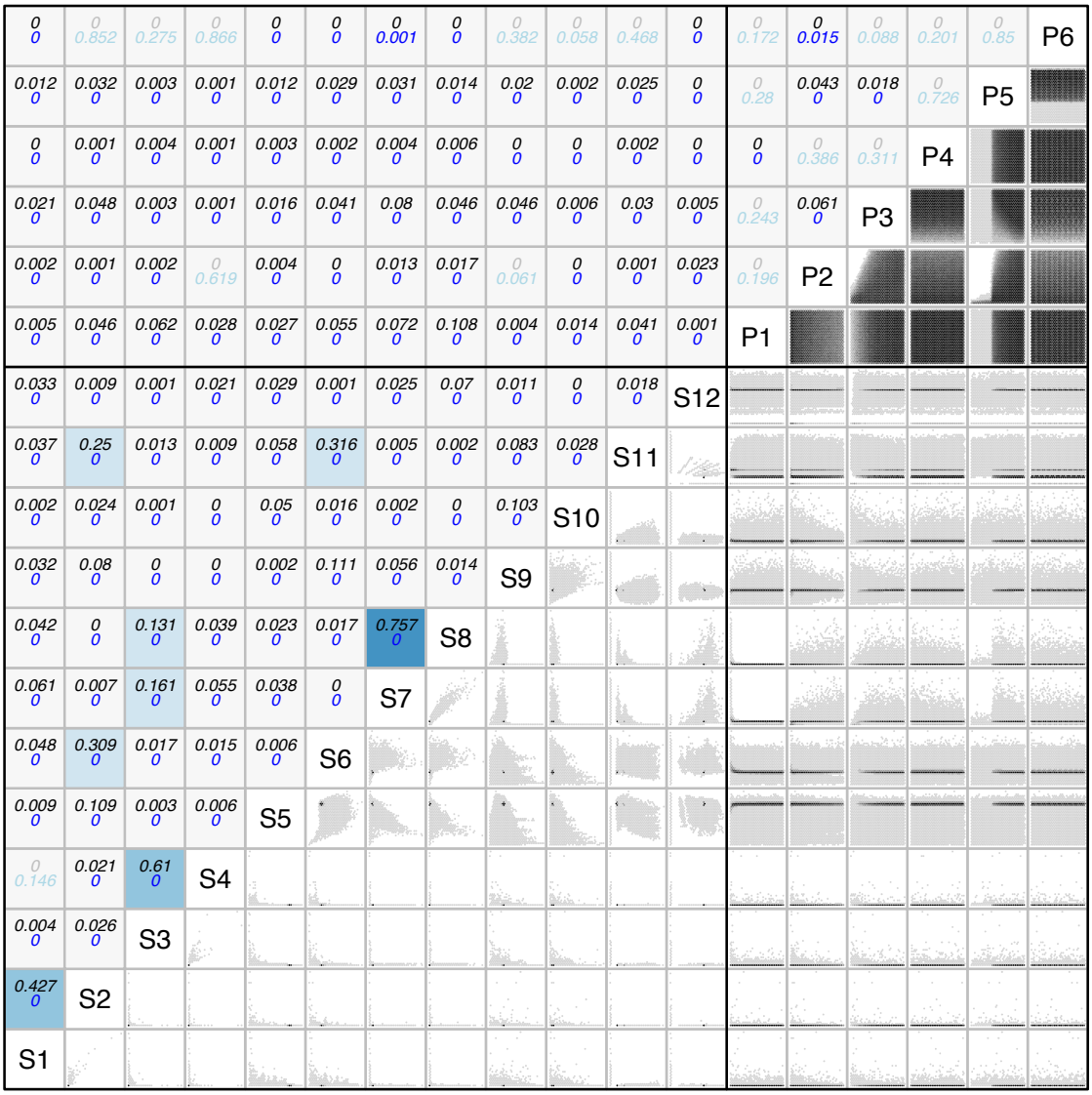


Figure A.9: Binned scatterplots (below diagonal) and correlation values (above diagonal; black: Spearman's rank correlation coefficient R^2, blue: associated p-value) of pairwise combinations of all summary statistics and model parameters using 400,000 successful simulations for the NMCI DIS B-A model.

The cells in each pairwise scatterplot (below diagonal) are shaded according to the number of data points they contain, with darker cells containing a larger number of data points than lighter cells. Pairwise panels giving the correlation values (above diagonal) are shaded according to R^2 value and the text greyed out for those that are not significant (p-value ≥ 0.05). P1: probability of cultural mutation; S2: probability of fission / extinction; P3: migration distance (km); P4: total number of items in each group's cultural repertoire; P5: maximum number of groups; P6: interaction radius (km); S1: mean(SI:ornaments); S2: var(SI:ornaments); S3: mean(SI:sites); S4: var(SI:sites); S5: mean(MI:ornaments); S6: var(MI:ornaments); S7: mean(MI:sites); S8: var(MI:sites); S9: mean(DR); S10: var(DR); S11: MAD:ornaments; S12: MAD:sites.

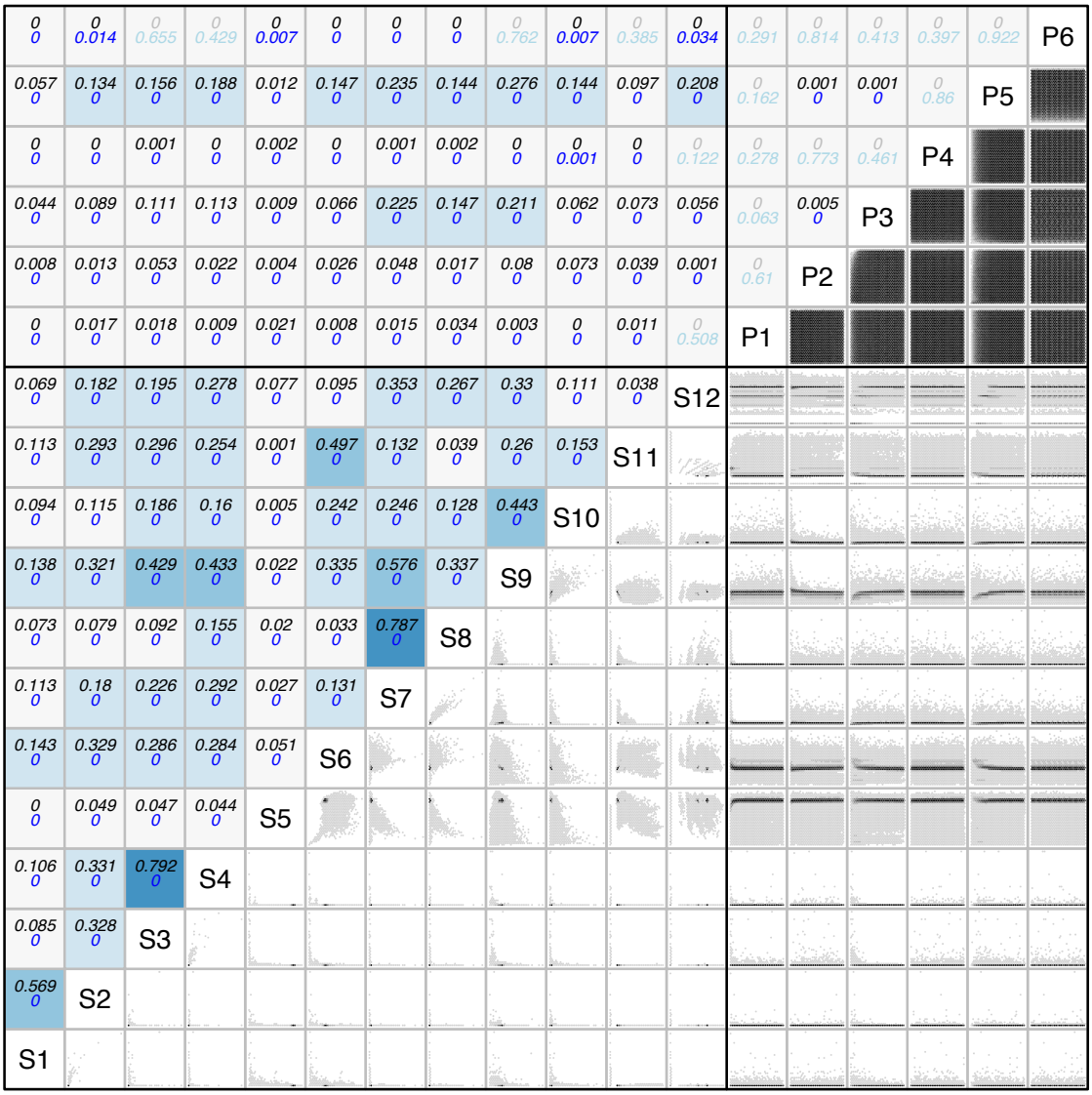
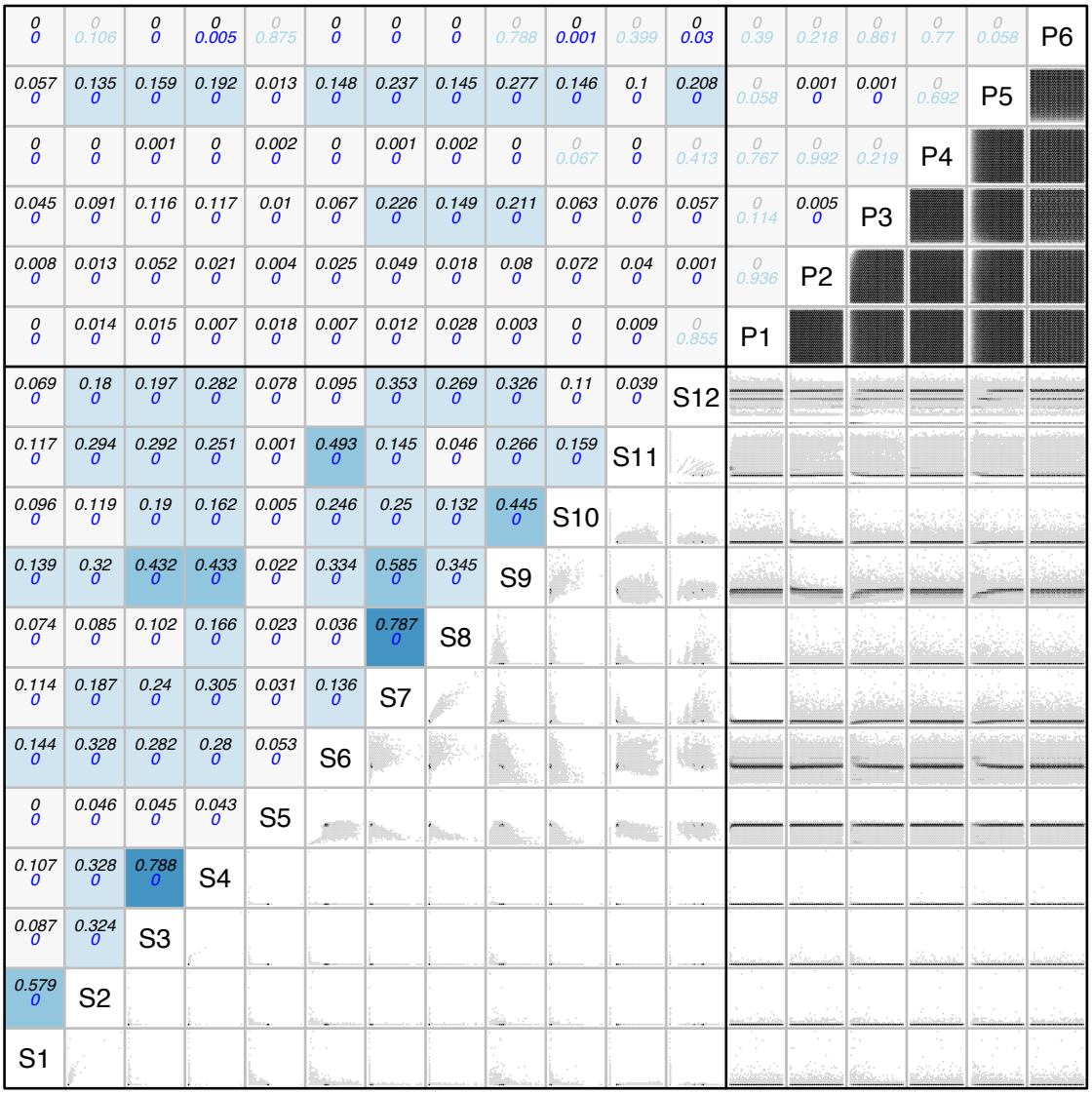


Figure A.10: Binned scatterplots (below diagonal) and correlation values (above diagonal; black: Spearman's rank correlation coefficient R^2, blue: associated p-value) of pairwise combinations of all summary statistics and model parameters using 400,000 successful simulations for the CD DIS B-A model.

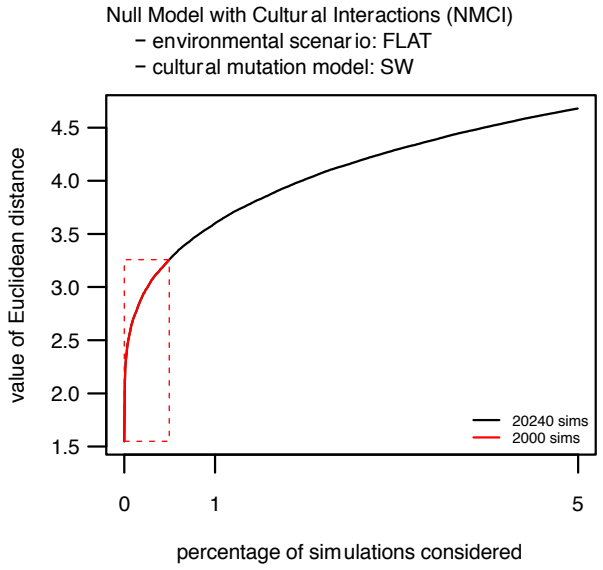
The cells in each pairwise scatterplot (below diagonal) are shaded according to the number of data points they contain, with darker cells containing a larger number of data points than lighter cells. Pairwise panels giving the correlation values (above diagonal) are shaded according to R^2 value and the text greyed out for those that are not significant (p-value ≥ 0.05). P1: probability of cultural mutation; S2: probability of fission / extinction; P3: migration distance (km); P4: total number of items in each group's cultural repertoire; P5: maximum number of groups; P6: interaction radius (km); S1: mean(SI:ornaments); S2: var(SI:ornaments); S3: mean(SI:sites); S4: var(SI:sites); S5: mean(MI:ornaments); S6: var(MI:ornaments); S7: mean(MI:sites); S8: var(MI:sites); S9: mean(DR); S10: var(DR); S11: MAD:ornaments; S12: MAD:sites.

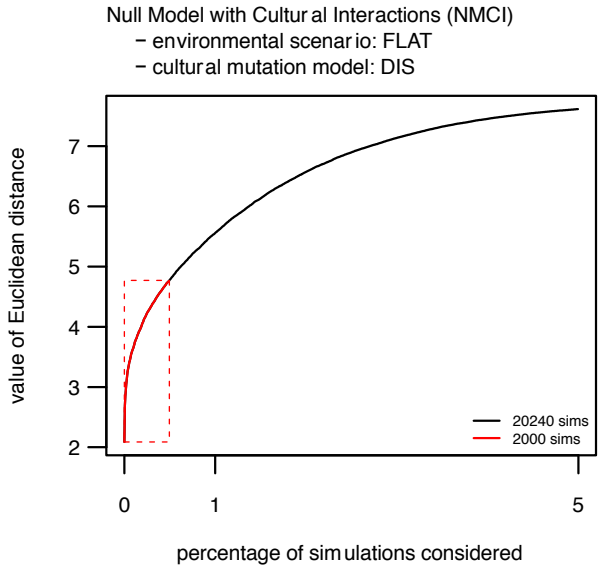


A.2 Parameter Estimation Analysis from Chapter 3 Section 3.1.3 performed with Different Threshold Values

A.2.1 Threshold: 0.5% (i.e. closest 2,000 of 404,808 simulations)

Figure A.11: Ranked Euclidean distance values of the best 20,240 simulations (i.e. the closest 5% of 404,808 simulations) for each of the two models plotted in black, with the 2,000 retained simulations (i.e. closest ~0.5% of 404,808 simulations) used for estimating the posterior parameter distributions of parameters highlighted in red. The two panels correspond to the two models of interest: NMCI SW FLAT (top panel) and NMCI DIS FLAT (bottom panel).





Estimated posterior density distributions of the demographic and evolutionary parameters of interest for the two NMCI models, calculated on the 2,000 retained simulations (i.e. closest ~0.5% of 404,808 simulations) for each model. The boundaries of the equal-tailed 95% credible intervals (i.e. the upper and lower 2.5%) of each distribution are indicated by shading; these are also summarised in Table A.1. The solid and dashed grey lines represent the prior and extinct (i.e. those simulations in which all groups became extinct prior to the end of the simulation) density distributions of parameters, respectively.

Figure A.12: NMCI SW FLAT model.

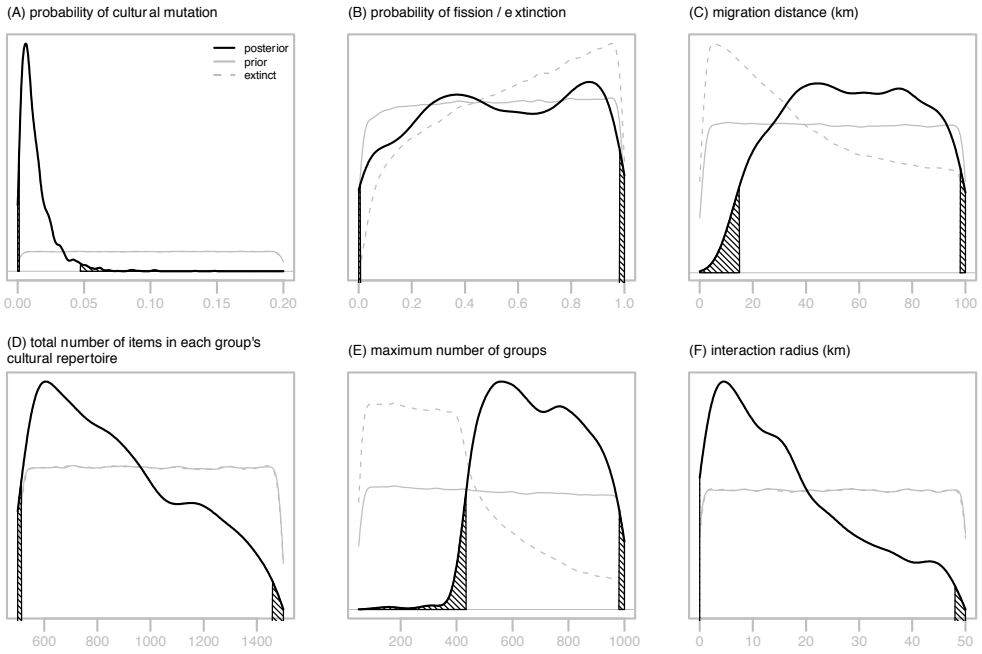


Figure A.13: NMCI DIS FLAT model.

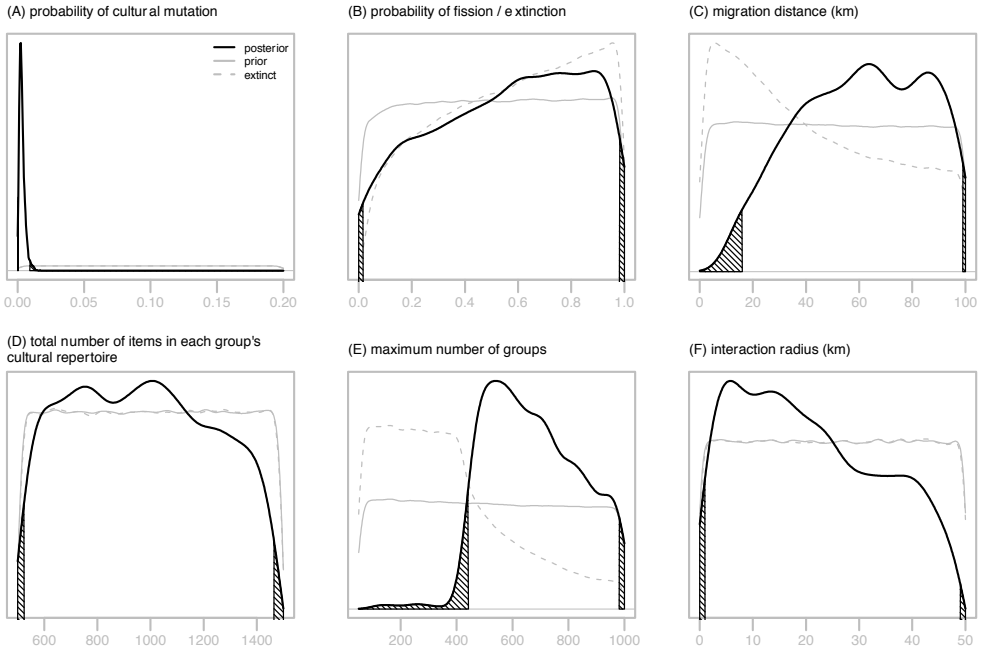
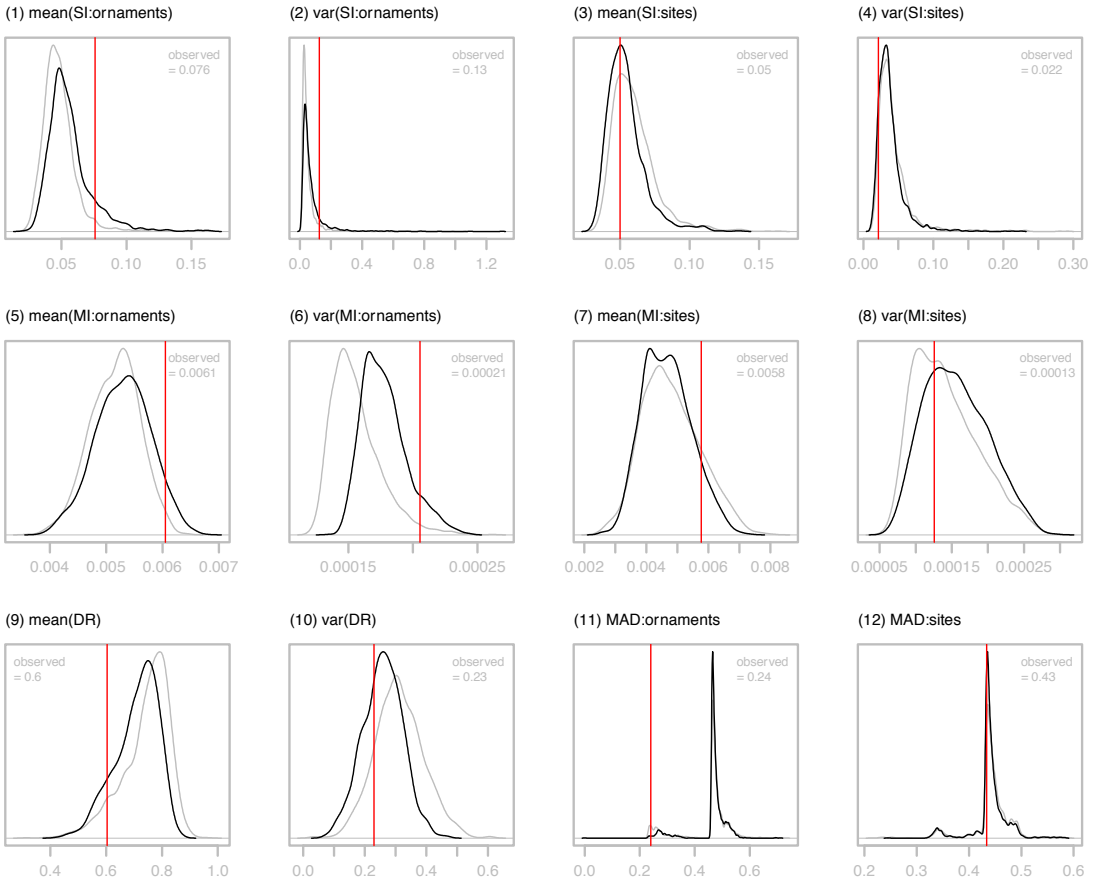


Table A.1: Prior ranges and posterior estimates of parameters for the two versions of interest of the Null Model with Cultural Interactions. For each model, the posterior parameter ranges are calculated on the 2,000 retained simulations (i.e. closest ~0.5% of 404,808 simulations) and expressed by giving the mode, 2.5% and 97.5% quantiles, expressed to 4 decimal places. The letter in the far left column corresponds to the panels in Figure A.12 (NMCI SW FLAT) and Figure A.13 (NMCI DIS FLAT).

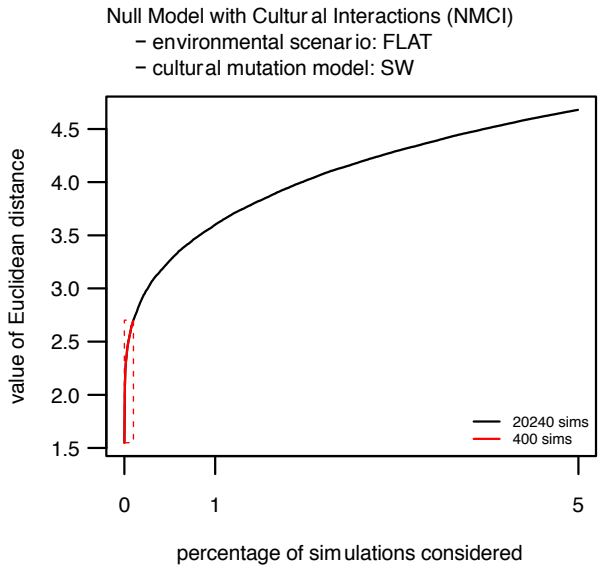
	Parameter	Prior Range		SW FLAT			DIS FLAT		
				Posterior Estimate					
		minimum	maximum	mode	2.5% quantile	97.5% quantile	mode	2.5% quantile	97.5% quantile
(A)	p_{mut}: probability of cultural mutation	0	0.2	0.0059	0.0013	0.0472	0.0023	0.0005	0.0091
(B)	$p_{f/e}$: probability of fission / extinction	0	1	0.8669	0.0073	0.9808	0.8865	0.0168	0.9805
(C)	d_{mig}: migration distance (km)	1	100	44.4427	15	98	63.7965	16	99
(D)	N_{items}: total number of items in each group's cultural repertoire	500	1500	606	515	1457	1007	525	1464
(E)	N_{groups}: maximum number of groups	50	1000	559	434	980	540	442	981
(F)	d_{int}: interaction radius (km)	0	50	4.5010	0	48	5.7730	1	49

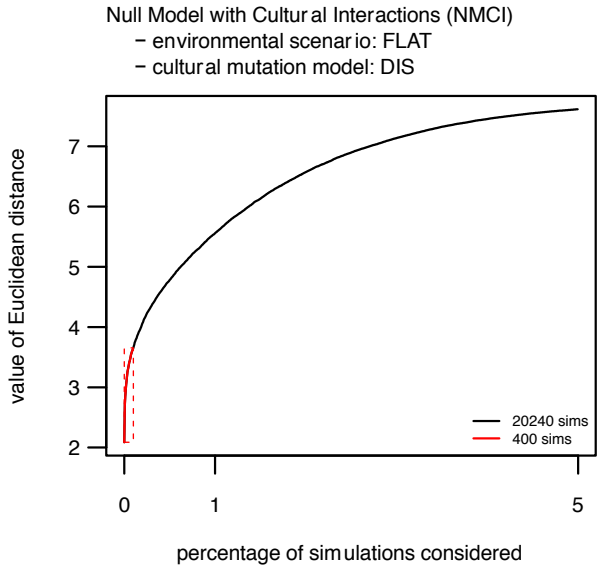
Figure A.14: Distributions of the 12 summary statistic values in the 2,000 retained simulations (i.e. closest ~0.5% of 404,808 simulations) for the NMCI SW FLAT model (black lines) and NMCI DIS FLAT model (grey lines). The title of each panel corresponds to the summary statistic as discussed in section 2.2.1. The red vertical line indicates the target value of each summary statistic (i.e. the value of that statistic calculated from the observed data).



A.2.2 Threshold: 0.1% (i.e. closest 400 of 404,808 simulations)

Figure A.15: Ranked Euclidean distance values of the best 20,240 simulations (i.e. the closest 5% of 404,808 simulations) for each of the two models plotted in black, with the 400 retained simulations (i.e. closest ~0.1% of 404,808 simulations) used for estimating the posterior parameter distributions of parameters highlighted in red. The two panels correspond to the two models of interest: NMCI SW FLAT (top panel) and NMCI DIS FLAT (bottom panel).





Estimated posterior density distributions of the demographic and evolutionary parameters of interest for the two NMCI models, calculated on the 400 retained simulations (i.e. closest ~0.1% of 404,808 simulations) for each model. The boundaries of the equal-tailed 95% credible intervals (i.e. the upper and lower 2.5%) of each distribution are indicated by shading; these are also summarised in Table A.2. The solid and dashed grey lines represent the prior and extinct (i.e. those simulations in which all groups became extinct prior to the end of the simulation) density distributions of parameters, respectively.

Figure A.16: NMCI SW FLAT model.

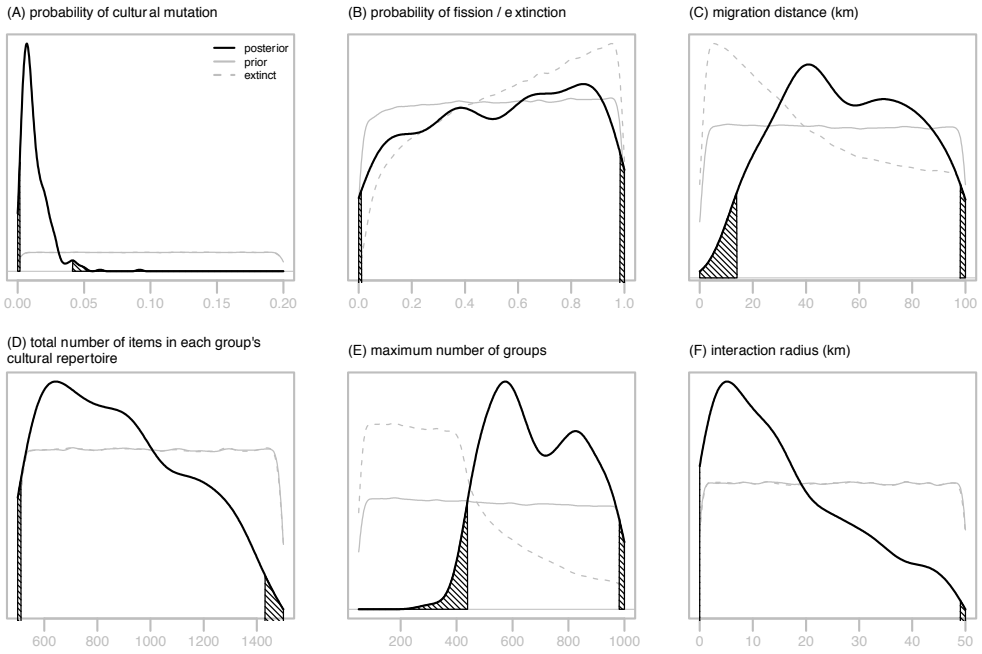


Figure A.17: NMCI DIS FLAT model.

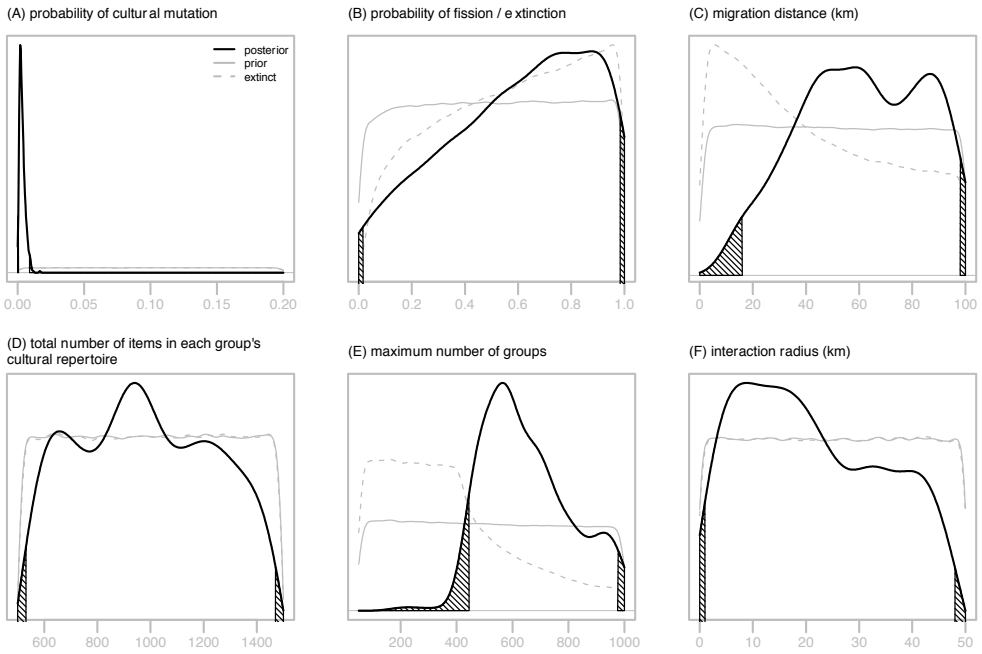
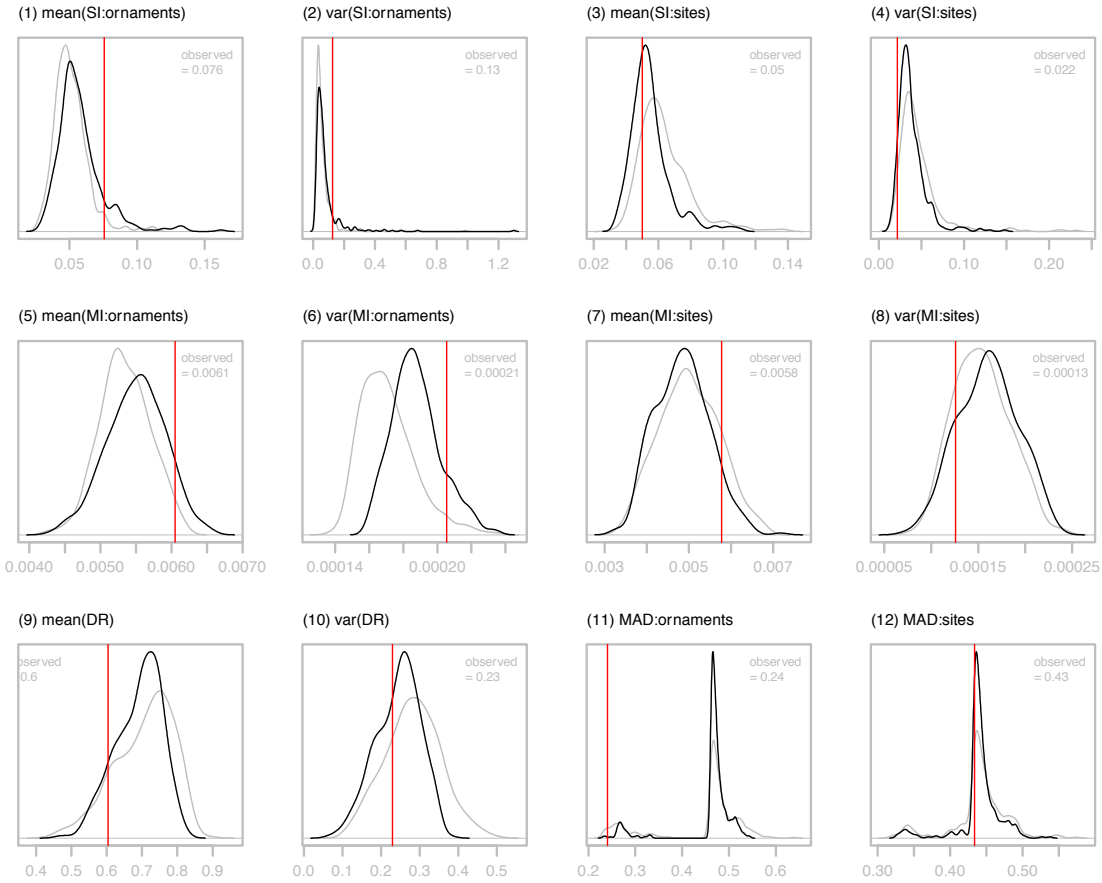


Table A.2: Prior ranges and posterior estimates of parameters for the two versions of interest of the Null Model with Cultural Interactions. For each model, the posterior parameter ranges are calculated on the 400 retained simulations (i.e. closest ~0.1% of 404,808 simulations) and expressed by giving the mode, 2.5% and 97.5% quantiles, expressed to 4 decimal places. The letter in the far left column corresponds to the panels in Figure A.16 (NMCI SW FLAT) and Figure A.17 (NMCI DIS FLAT).

	Parameter	Prior Range		SW FLAT			DIS FLAT		
				Posterior Estimate					
		minimum	maximum	mode	2.5% quantile	97.5% quantile	mode	2.5% quantile	97.5% quantile
(A)	p_{mut}: probability of cultural mutation	0	0.2	0.0070	0.0017	0.0412	0.0020	0.0007	0.0090
(B)	$p_{f/e}$: probability of fission / extinction	0	1	0.8454	0.0115	0.9824	0.8767	0.0168	0.9840
(C)	d_{mig}: migration distance (km)	1	100	40.9002	14	98	58.5127	16	98
(D)	N_{items}: total number of items in each group's cultural repertoire	500	1500	643	513	1430	940	532	1469
(E)	N_{groups}: maximum number of groups	50	1000	574	439	981	565	445	976
(F)	d_{int}: interaction radius (km)	0	50	5.1859	0	49	8.8063	1	48

Figure A.18: Distributions of the 12 summary statistic values in the 400 retained simulations (i.e. closest ~0.1% of 404,808 simulations) for the NMCI SW FLAT model (black lines) and NMCI DIS FLAT model (grey lines). The title of each panel corresponds to the summary statistic as discussed in section 2.2.1. The red vertical line indicates the target value of each summary statistic (i.e. the value of that statistic calculated from the observed data).



A.3 Correlation Plots for Models Described in Chapter 4

Figure A.19: Binned scatterplots (below diagonal) and correlation values (above diagonal; black: Spearman's rank correlation coefficient R^2, blue: associated p-value) of pairwise combinations of all summary statistics and model parameters using 1,000,000 successful simulations for the NMCI SW FLAT model.

The cells in each pairwise scatterplot (below diagonal) are shaded according to the number of data points they contain, with darker cells containing a larger number of data points than lighter cells. Pairwise panels giving the correlation values (above diagonal) are shaded according to R^2 value and the text greyed out for those that are not significant (p-value ≥ 0.05). P1: probability of cultural mutation; S2: probability of fission / extinction; P3: migration distance (km); P4: total number of items in each group's cultural repertoire; P5: maximum number of groups; P6: interaction radius (km); S1: mean(SI:ornaments); S2: var(SI:ornaments); S3: mean(SI:sites); S4: var(SI:sites); S5: mean(MI:ornaments); S6: var(MI:ornaments); S7: mean(MI:sites); S8: var(MI:sites); S9: mean(DR); S10: var(DR); S11: MAD:ornaments; S12: MAD:sites.

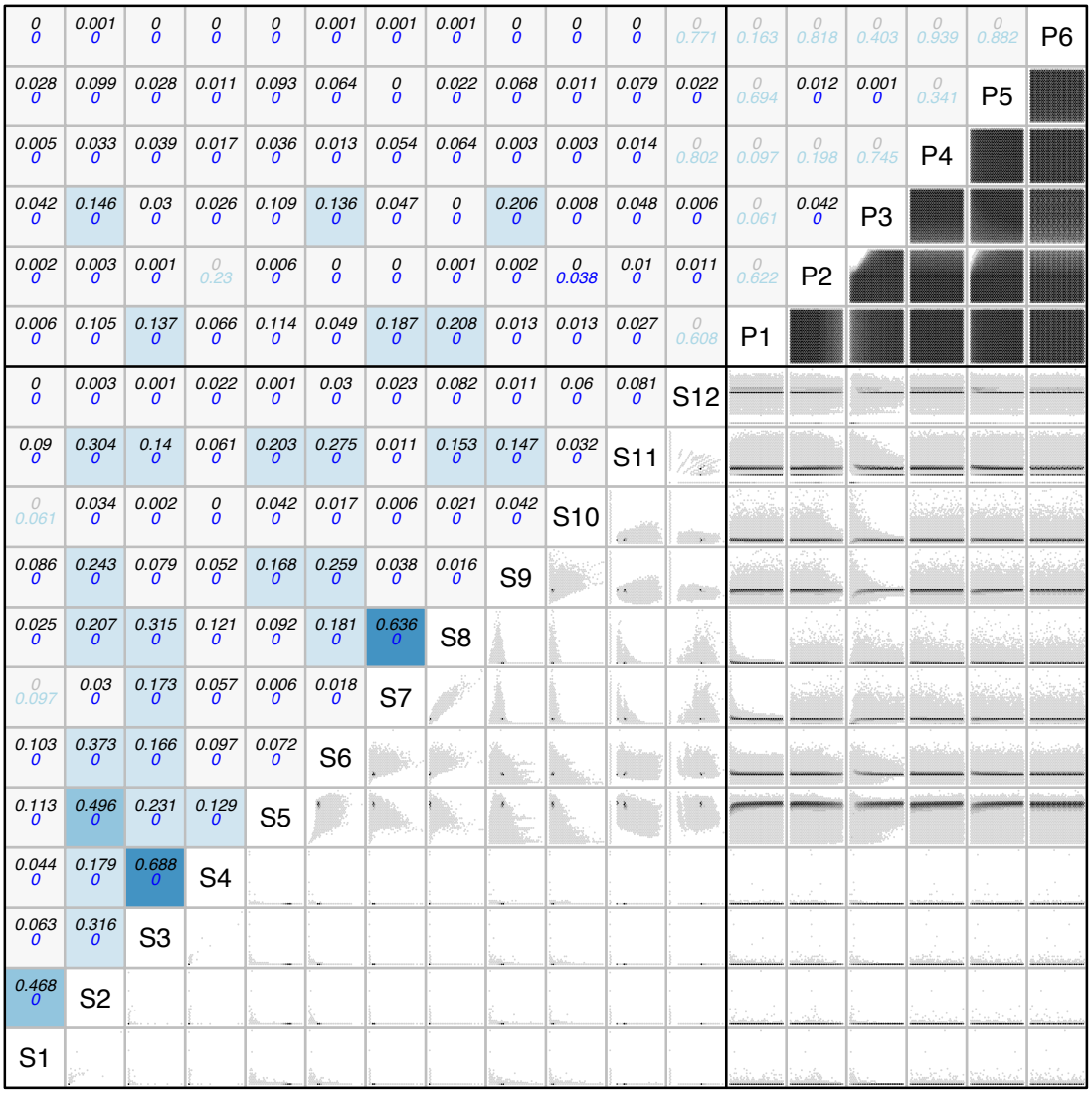


Figure A.20: Binned scatterplots (below diagonal) and correlation values (above diagonal; black: Spearman's rank correlation coefficient R^2, blue: associated p-value) of pairwise combinations of all summary statistics and model parameters using 1,000,000 successful simulations for the CD SW FLAT model.

The cells in each pairwise scatterplot (below diagonal) are shaded according to the number of data points they contain, with darker cells containing a larger number of data points than lighter cells. Pairwise panels giving the correlation values (above diagonal) are shaded according to R^2 value and the text greyed out for those that are not significant (p-value ≥ 0.05). P1: probability of cultural mutation; S2: probability of fission / extinction; P3: migration distance (km); P4: total number of items in each group's cultural repertoire; P5: maximum number of groups; P6: interaction radius (km); S1: mean(SI:ornaments); S2: var(SI:ornaments); S3: mean(SI:sites); S4: var(SI:sites); S5: mean(MI:ornaments); S6: var(MI:ornaments); S7: mean(MI:sites); S8: var(MI:sites); S9: mean(DR); S10: var(DR); S11: MAD:ornaments; S12: MAD:sites.

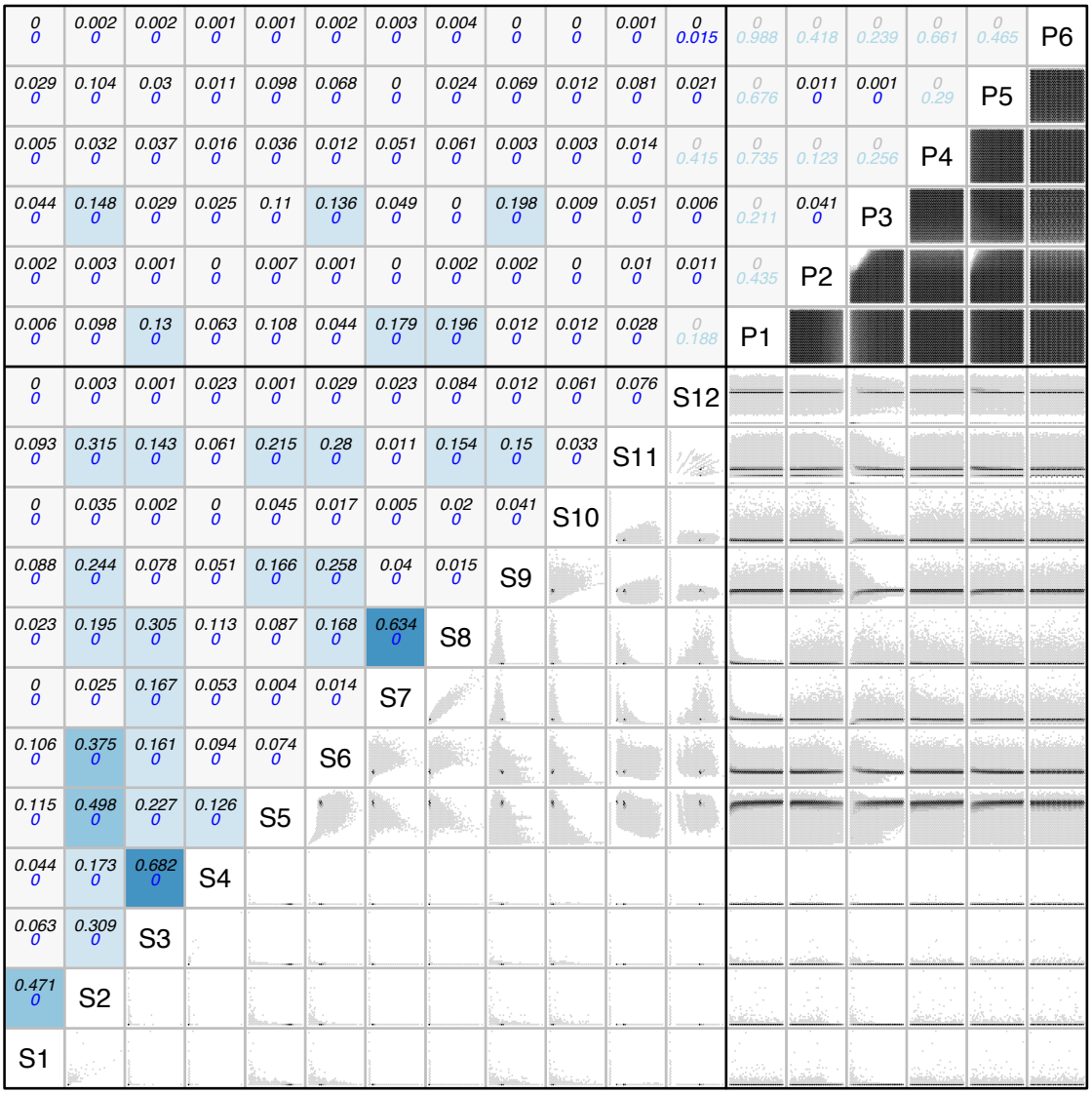


Figure A.21: Binned scatterplots (below diagonal) and correlation values (above diagonal; black: Spearman's rank correlation coefficient R^2, blue: associated p-value) of pairwise combinations of all summary statistics and model parameters using 1,000,000 successful simulations for the NMCI SW NPPBanks model.

The cells in each pairwise scatterplot (below diagonal) are shaded according to the number of data points they contain, with darker cells containing a larger number of data points than lighter cells. Pairwise panels giving the correlation values (above diagonal) are shaded according to R^2 value and the text greyed out for those that are not significant (p-value ≥ 0.05). P1: probability of cultural mutation; S2: probability of fission / extinction; P3: migration distance (km); P4: total number of items in each group's cultural repertoire; P5: maximum number of groups; P6: interaction radius (km); S1: mean(SI:ornaments); S2: var(SI:ornaments); S3: mean(SI:sites); S4: var(SI:sites); S5: mean(MI:ornaments); S6: var(MI:ornaments); S7: mean(MI:sites); S8: var(MI:sites); S9: mean(DR); S10: var(DR); S11: MAD:ornaments; S12: MAD:sites.

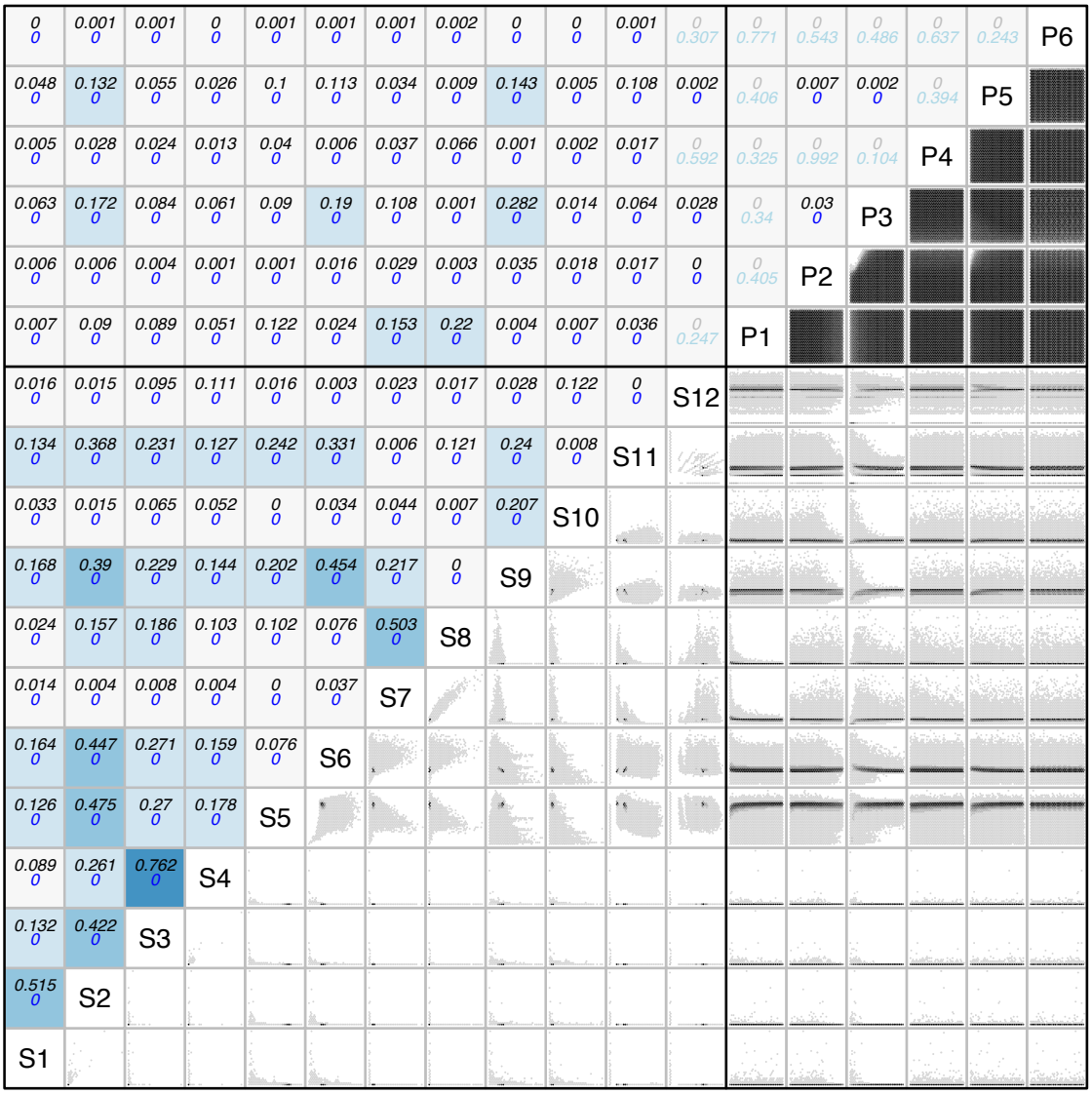


Figure A.22: Binned scatterplots (below diagonal) and correlation values (above diagonal; black: Spearman's rank correlation coefficient R^2, blue: associated p-value) of pairwise combinations of all summary statistics and model parameters using 1,000,000 successful simulations for the CD SW NPPBanks model.

The cells in each pairwise scatterplot (below diagonal) are shaded according to the number of data points they contain, with darker cells containing a larger number of data points than lighter cells. Pairwise panels giving the correlation values (above diagonal) are shaded according to R^2 value and the text greyed out for those that are not significant (p-value ≥ 0.05). P1: probability of cultural mutation; S2: probability of fission / extinction; P3: migration distance (km); P4: total number of items in each group's cultural repertoire; P5: maximum number of groups; P6: interaction radius (km); S1: mean(SI:ornaments); S2: var(SI:ornaments); S3: mean(SI:sites); S4: var(SI:sites); S5: mean(MI:ornaments); S6: var(MI:ornaments); S7: mean(MI:sites); S8: var(MI:sites); S9: mean(DR); S10: var(DR); S11: MAD:ornaments; S12: MAD:sites.

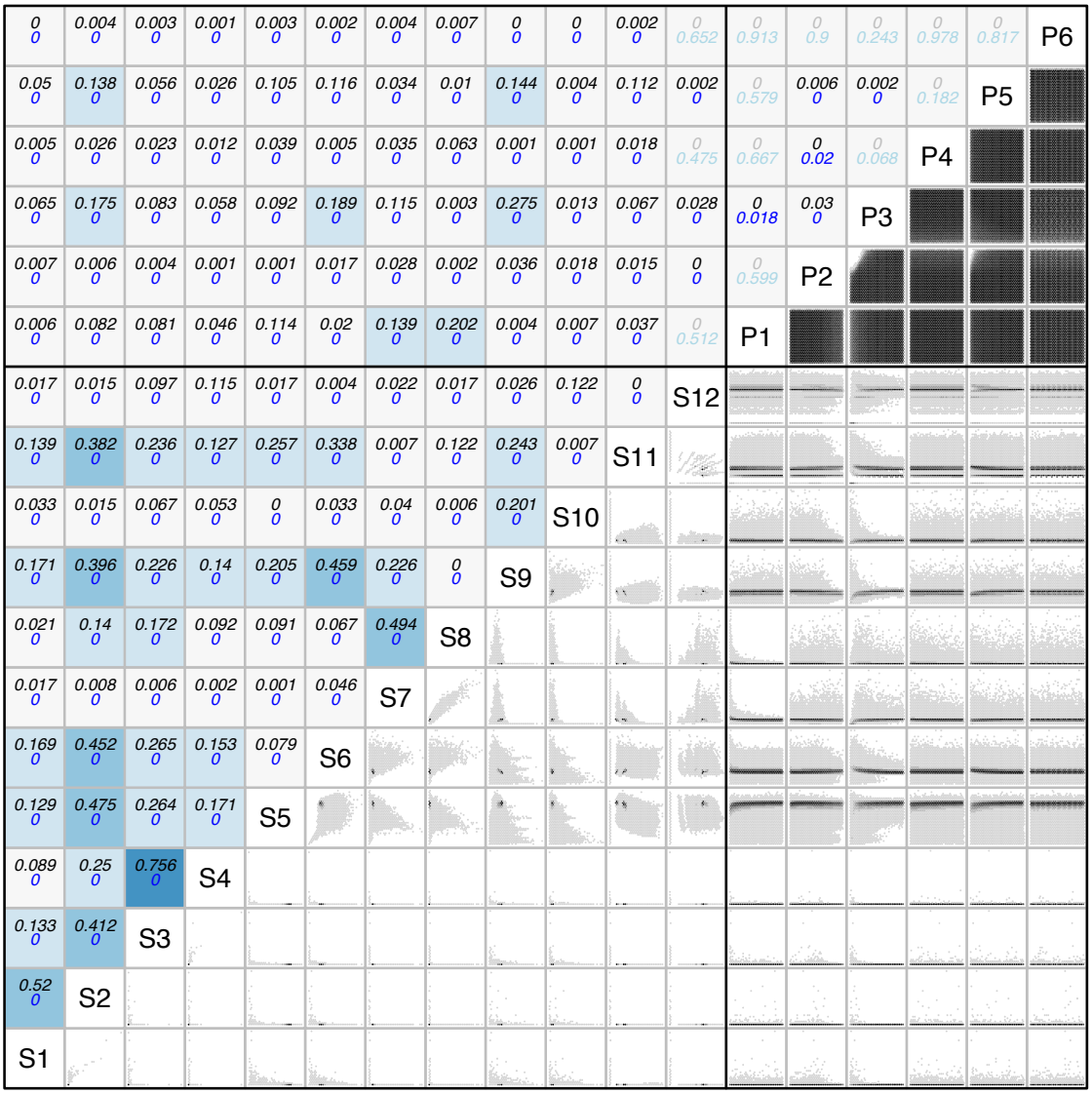


Figure A.23: Binned scatterplots (below diagonal) and correlation values (above diagonal; black: Spearman's rank correlation coefficient R^2, blue: associated p-value) of pairwise combinations of all summary statistics and model parameters using 1,000,000 successful simulations for the NMCI SW NPPSingarayer model.

The cells in each pairwise scatterplot (below diagonal) are shaded according to the number of data points they contain, with darker cells containing a larger number of data points than lighter cells. Pairwise panels giving the correlation values (above diagonal) are shaded according to R^2 value and the text greyed out for those that are not significant (p-value ≥ 0.05). P1: probability of cultural mutation; S2: probability of fission / extinction; P3: migration distance (km); P4: total number of items in each group's cultural repertoire; P5: maximum number of groups; P6: interaction radius (km); S1: mean(SI:ornaments); S2: var(SI:ornaments); S3: mean(SI:sites); S4: var(SI:sites); S5: mean(MI:ornaments); S6: var(MI:ornaments); S7: mean(MI:sites); S8: var(MI:sites); S9: mean(DR); S10: var(DR); S11: MAD:ornaments; S12: MAD:sites.

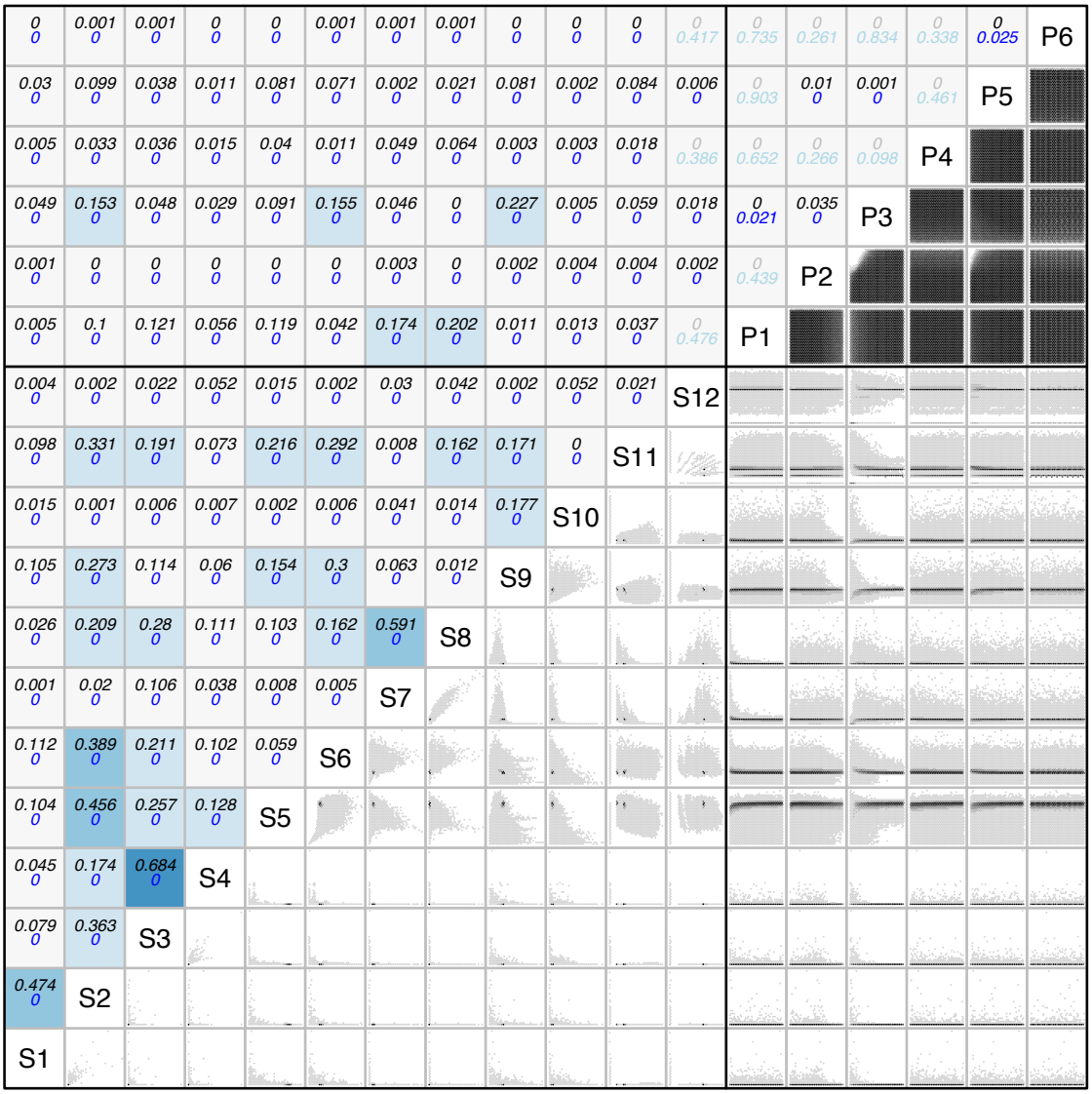


Figure A.24: Binned scatterplots (below diagonal) and correlation values (above diagonal; black: Spearman's rank correlation coefficient R^2, blue: associated p-value) of pairwise combinations of all summary statistics and model parameters using 1,000,000 successful simulations for the CD SW NPPSingarayer model.

The cells in each pairwise scatterplot (below diagonal) are shaded according to the number of data points they contain, with darker cells containing a larger number of data points than lighter cells. Pairwise panels giving the correlation values (above diagonal) are shaded according to R^2 value and the text greyed out for those that are not significant (p-value ≥ 0.05). P1: probability of cultural mutation; S2: probability of fission / extinction; P3: migration distance (km); P4: total number of items in each group's cultural repertoire; P5: maximum number of groups; P6: interaction radius (km); S1: mean(SI:ornaments); S2: var(SI:ornaments); S3: mean(SI:sites); S4: var(SI:sites); S5: mean(MI:ornaments); S6: var(MI:ornaments); S7: mean(MI:sites); S8: var(MI:sites); S9: mean(DR); S10: var(DR); S11: MAD:ornaments; S12: MAD:sites.

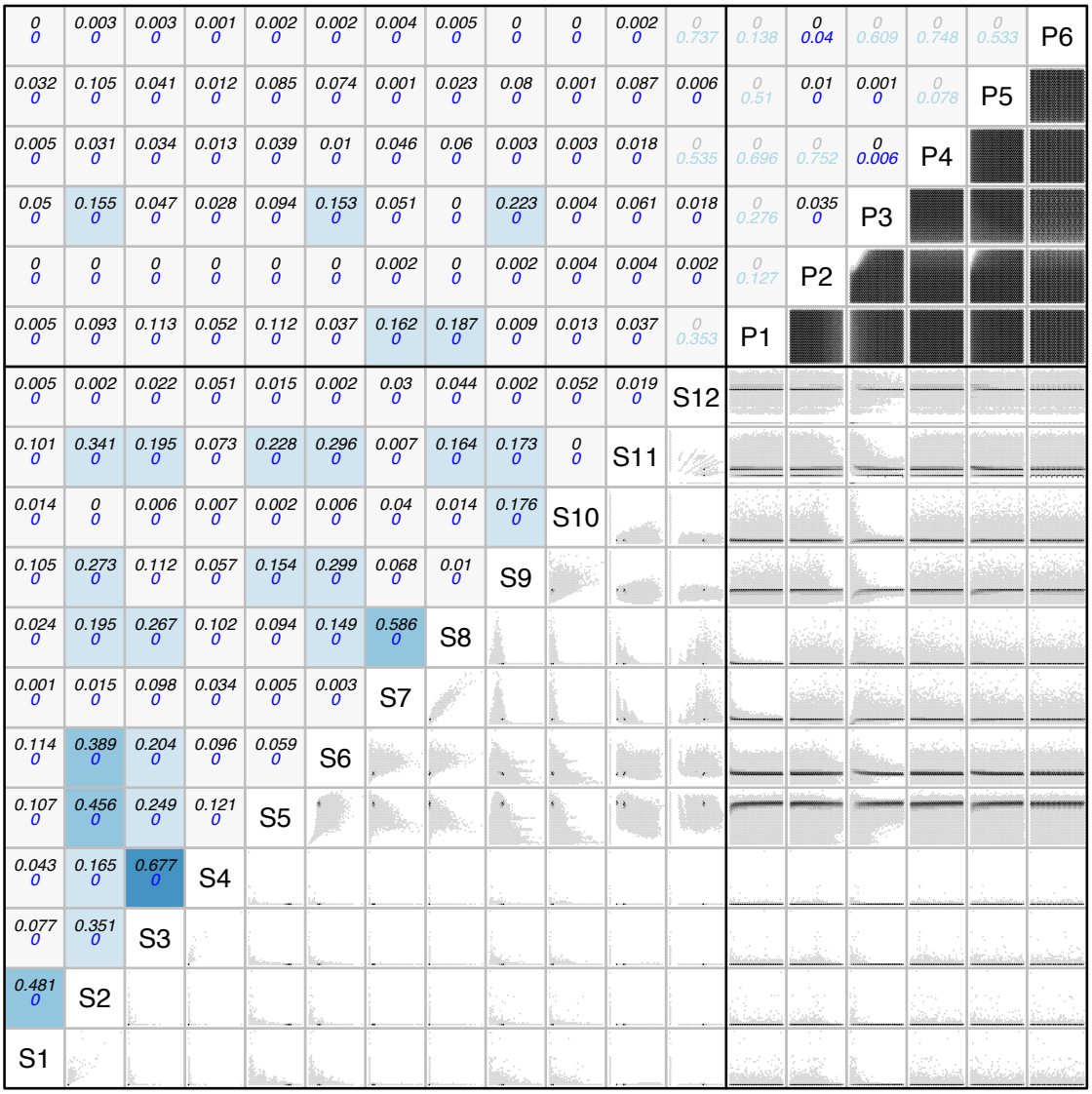


Figure A.25: Binned scatterplots (below diagonal) and correlation values (above diagonal; black: Spearman's rank correlation coefficient R^2, blue: associated p-value) of pairwise combinations of all summary statistics and model parameters using 1,000,000 successful simulations for the NMCI DIS FLAT model.

The cells in each pairwise scatterplot (below diagonal) are shaded according to the number of data points they contain, with darker cells containing a larger number of data points than lighter cells. Pairwise panels giving the correlation values (above diagonal) are shaded according to R^2 value and the text greyed out for those that are not significant (p-value ≥ 0.05). P1: probability of cultural mutation; S2: probability of fission / extinction; P3: migration distance (km); P4: total number of items in each group's cultural repertoire; P5: maximum number of groups; P6: interaction radius (km); S1: mean(SI:ornaments); S2: var(SI:ornaments); S3: mean(SI:sites); S4: var(SI:sites); S5: mean(MI:ornaments); S6: var(MI:ornaments); S7: mean(MI:sites); S8: var(MI:sites); S9: mean(DR); S10: var(DR); S11: MAD:ornaments; S12: MAD:sites.

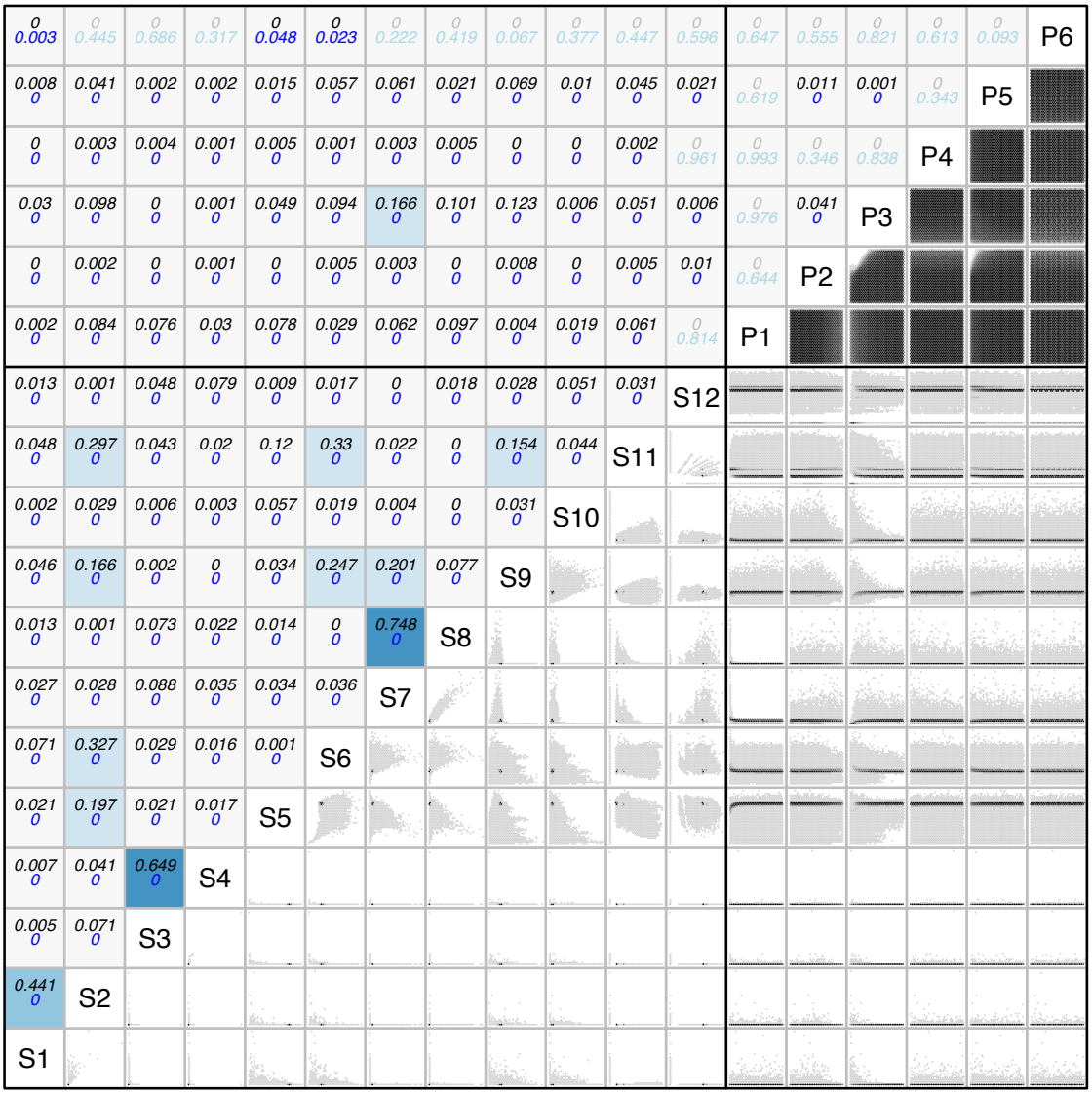


Figure A.26: Binned scatterplots (below diagonal) and correlation values (above diagonal; black: Spearman's rank correlation coefficient R^2, blue: associated p-value) of pairwise combinations of all summary statistics and model parameters using 1,000,000 successful simulations for the CD DIS FLAT model.

The cells in each pairwise scatterplot (below diagonal) are shaded according to the number of data points they contain, with darker cells containing a larger number of data points than lighter cells. Pairwise panels giving the correlation values (above diagonal) are shaded according to R^2 value and the text greyed out for those that are not significant (p-value ≥ 0.05). P1: probability of cultural mutation; S2: probability of fission / extinction; P3: migration distance (km); P4: total number of items in each group's cultural repertoire; P5: maximum number of groups; P6: interaction radius (km); S1: mean(SI:ornaments); S2: var(SI:ornaments); S3: mean(SI:sites); S4: var(SI:sites); S5: mean(MI:ornaments); S6: var(MI:ornaments); S7: mean(MI:sites); S8: var(MI:sites); S9: mean(DR); S10: var(DR); S11: MAD:ornaments; S12: MAD:sites.

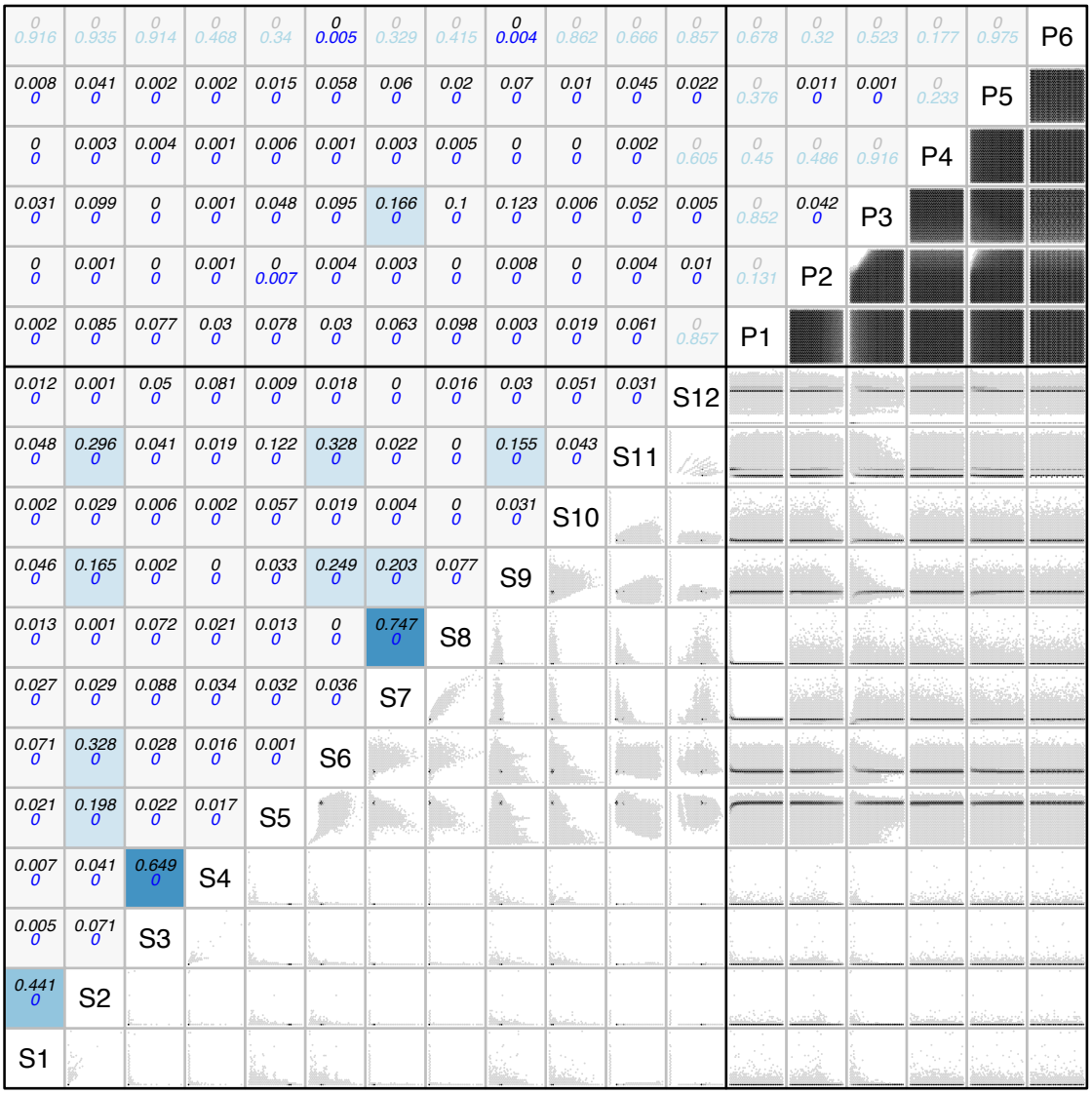


Figure A.27: Binned scatterplots (below diagonal) and correlation values (above diagonal; black: Spearman's rank correlation coefficient R^2, blue: associated p-value) of pairwise combinations of all summary statistics and model parameters using 1,000,000 successful simulations for the NMCI DIS NPPBanks model.

The cells in each pairwise scatterplot (below diagonal) are shaded according to the number of data points they contain, with darker cells containing a larger number of data points than lighter cells. Pairwise panels giving the correlation values (above diagonal) are shaded according to R^2 value and the text greyed out for those that are not significant (p-value ≥ 0.05). P1: probability of cultural mutation; S2: probability of fission / extinction; P3: migration distance (km); P4: total number of items in each group's cultural repertoire; P5: maximum number of groups; P6: interaction radius (km); S1: mean(SI:ornaments); S2: var(SI:ornaments); S3: mean(SI:sites); S4: var(SI:sites); S5: mean(MI:ornaments); S6: var(MI:ornaments); S7: mean(MI:sites); S8: var(MI:sites); S9: mean(DR); S10: var(DR); S11: MAD:ornaments; S12: MAD:sites.

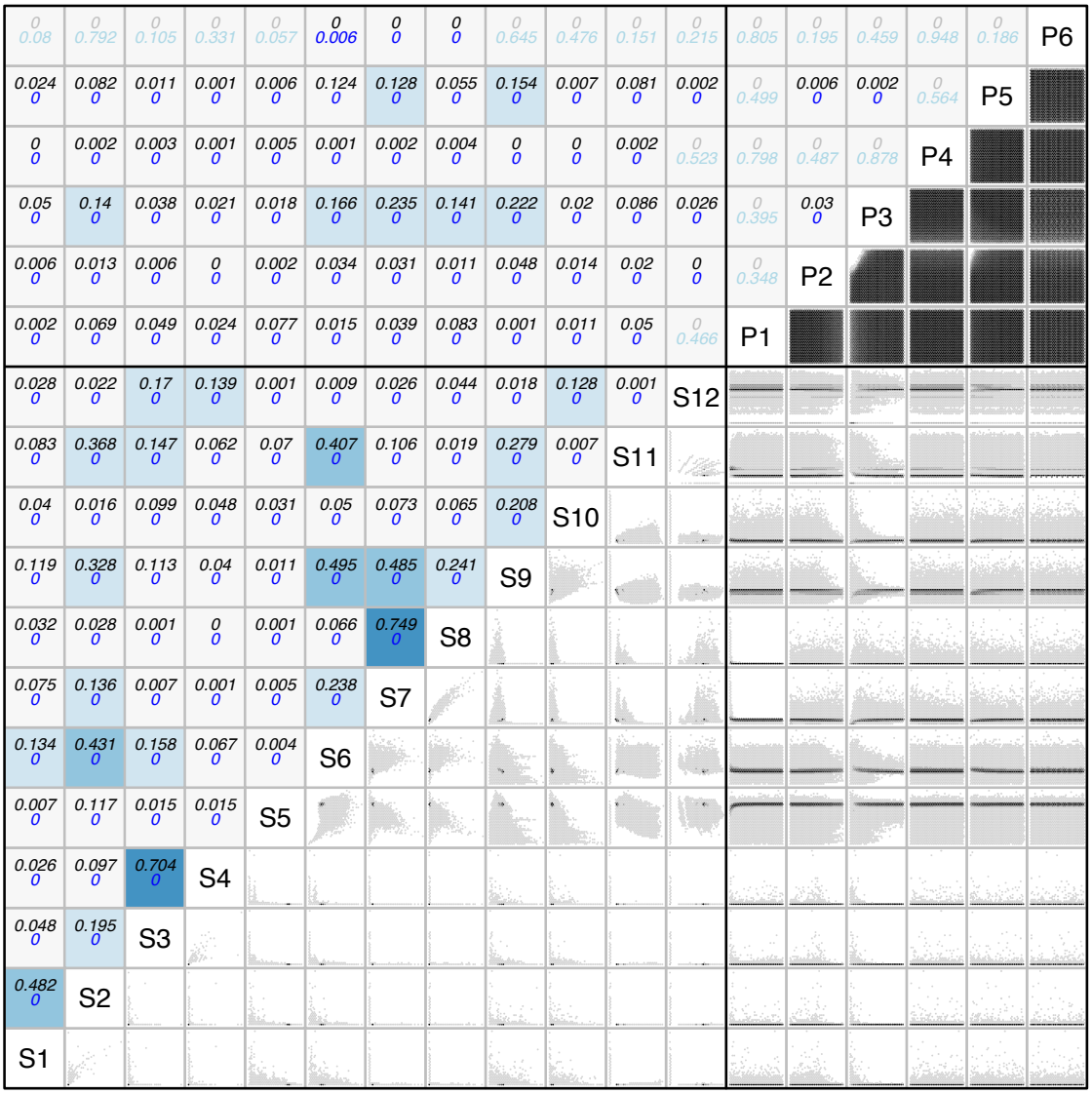


Figure A.28: Binned scatterplots (below diagonal) and correlation values (above diagonal; black: Spearman's rank correlation coefficient R^2, blue: associated p-value) of pairwise combinations of all summary statistics and model parameters using 1,000,000 successful simulations for the CD DIS NPPBanks model.

The cells in each pairwise scatterplot (below diagonal) are shaded according to the number of data points they contain, with darker cells containing a larger number of data points than lighter cells. Pairwise panels giving the correlation values (above diagonal) are shaded according to R^2 value and the text greyed out for those that are not significant (p-value ≥ 0.05). P1: probability of cultural mutation; S2: probability of fission / extinction; P3: migration distance (km); P4: total number of items in each group's cultural repertoire; P5: maximum number of groups; P6: interaction radius (km); S1: mean(SI:ornaments); S2: var(SI:ornaments); S3: mean(SI:sites); S4: var(SI:sites); S5: mean(MI:ornaments); S6: var(MI:ornaments); S7: mean(MI:sites); S8: var(MI:sites); S9: mean(DR); S10: var(DR); S11: MAD:ornaments; S12: MAD:sites.

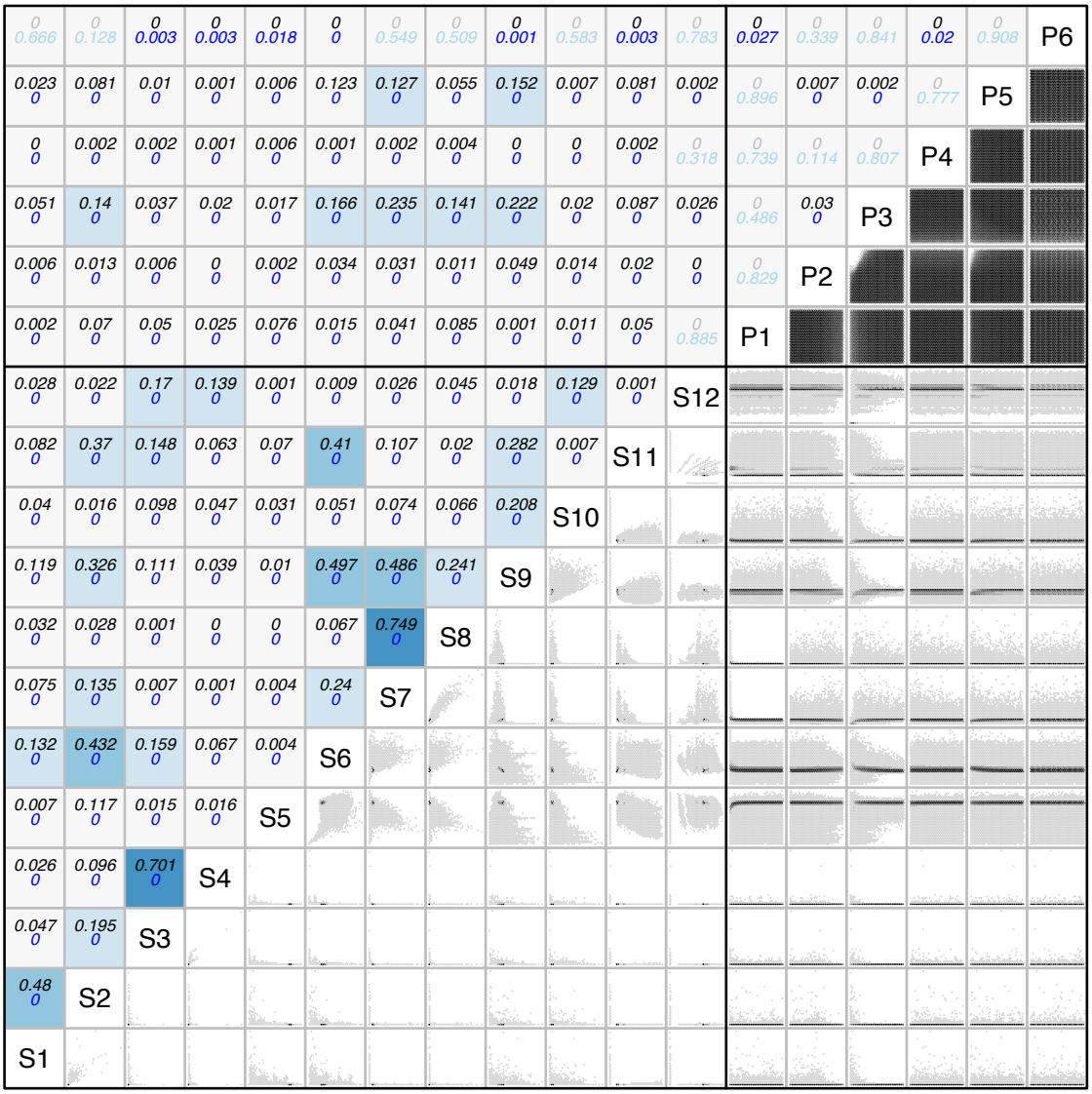


Figure A.29: Binned scatterplots (below diagonal) and correlation values (above diagonal; black: Spearman's rank correlation coefficient R^2, blue: associated p-value) of pairwise combinations of all summary statistics and model parameters using 1,000,000 successful simulations for the NMCI DIS NPPSingarayer model.

The cells in each pairwise scatterplot (below diagonal) are shaded according to the number of data points they contain, with darker cells containing a larger number of data points than lighter cells. Pairwise panels giving the correlation values (above diagonal) are shaded according to R^2 value and the text greyed out for those that are not significant (p-value ≥ 0.05). P1: probability of cultural mutation; S2: probability of fission / extinction; P3: migration distance (km); P4: total number of items in each group's cultural repertoire; P5: maximum number of groups; P6: interaction radius (km); S1: mean(SI:ornaments); S2: var(SI:ornaments); S3: mean(SI:sites); S4: var(SI:sites); S5: mean(MI:ornaments); S6: var(MI:ornaments); S7: mean(MI:sites); S8: var(MI:sites); S9: mean(DR); S10: var(DR); S11: MAD:ornaments; S12: MAD:sites.

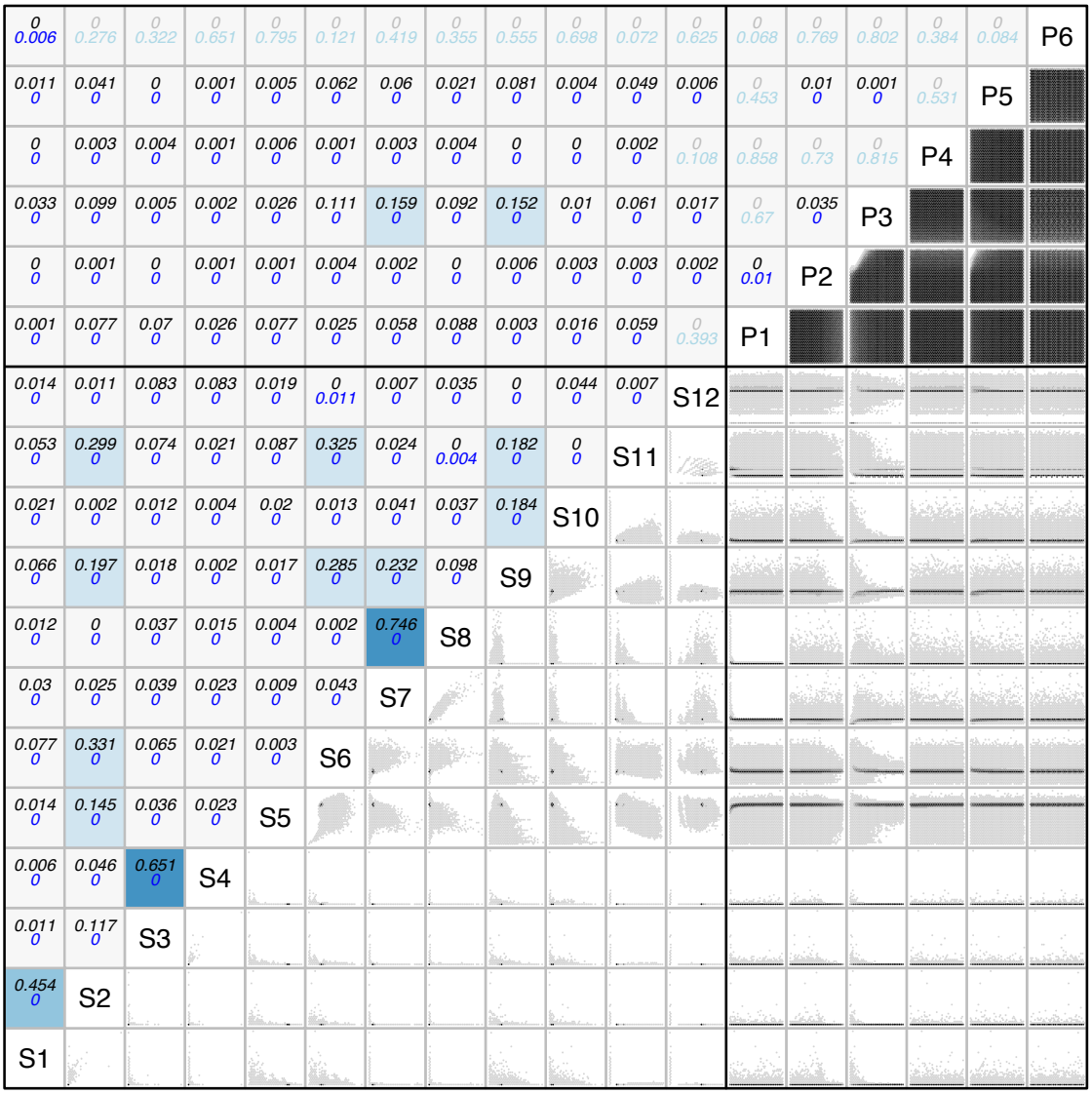
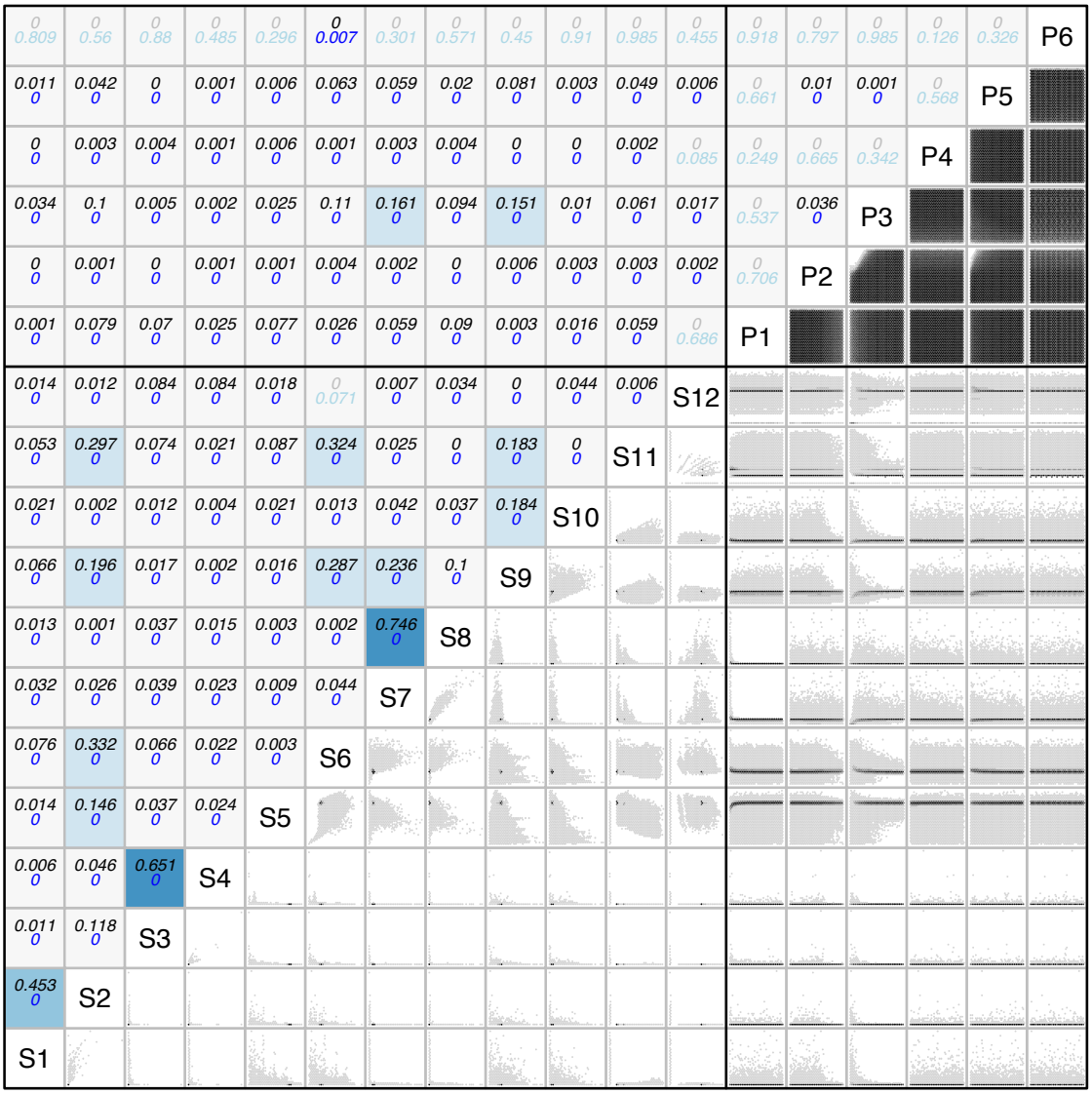


Figure A.30: Binned scatterplots (below diagonal) and correlation values (above diagonal; black: Spearman's rank correlation coefficient R^2, blue: associated p-value) of pairwise combinations of all summary statistics and model parameters using 1,000,000 successful simulations for the CD DIS NPPSingarayer model.

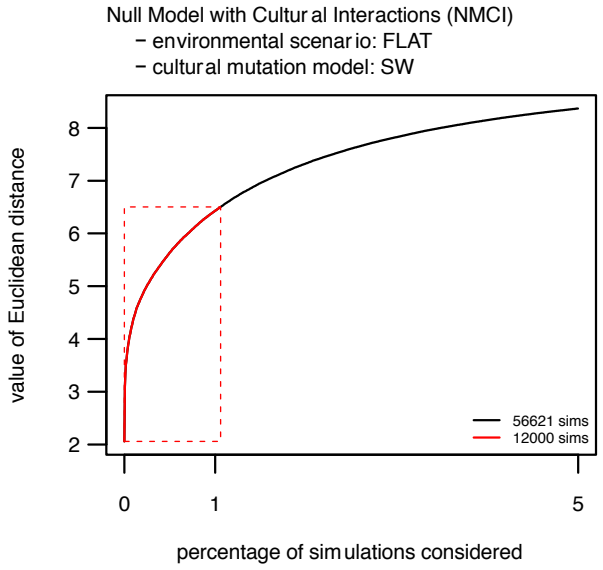
The cells in each pairwise scatterplot (below diagonal) are shaded according to the number of data points they contain, with darker cells containing a larger number of data points than lighter cells. Pairwise panels giving the correlation values (above diagonal) are shaded according to R^2 value and the text greyed out for those that are not significant (p-value ≥ 0.05). P1: probability of cultural mutation; S2: probability of fission / extinction; P3: migration distance (km); P4: total number of items in each group's cultural repertoire; P5: maximum number of groups; P6: interaction radius (km); S1: mean(SI:ornaments); S2: var(SI:ornaments); S3: mean(SI:sites); S4: var(SI:sites); S5: mean(MI:ornaments); S6: var(MI:ornaments); S7: mean(MI:sites); S8: var(MI:sites); S9: mean(DR); S10: var(DR); S11: MAD:ornaments; S12: MAD:sites.

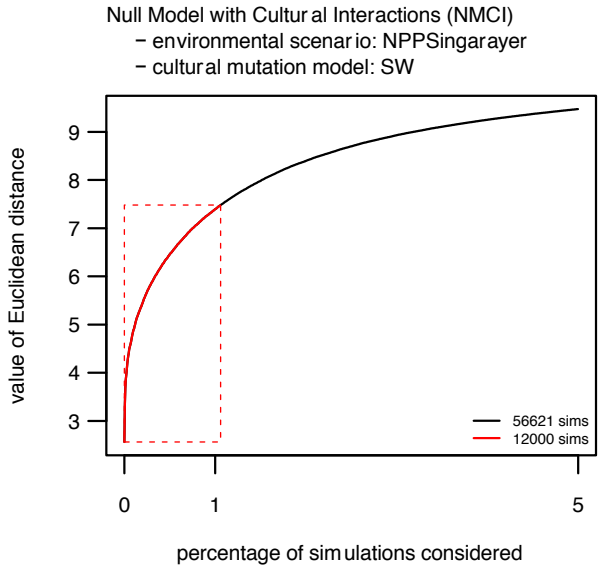


A.4 Parameter Estimation Analysis from Chapter 4 Section 4.2.3 performed with Different Threshold Values

A.4.1 Threshold: 1% (i.e. closest 12,000 of 1,132,411 simulations)

Figure A.31: Ranked Euclidean distance values of the best 56,621 simulations (i.e. the closest 5% of 1,132,411 simulations) for each of the two models plotted in black, with the 12,000 retained simulations (i.e. closest ~1% of 1,132,411 simulations) used for estimating the posterior parameter distributions of parameters highlighted in red. The two panels correspond to the two models of interest: NMCI SW FLAT (top panel) and NMCI SW NPPSingarayer (bottom panel).





Estimated posterior density distributions of the demographic and evolutionary parameters of interest for the two NMCI models, calculated on the 12,000 retained simulations (i.e. closest ~1% of 1,132,411 simulations) for each model. The boundaries of the equal-tailed 95% credible intervals (i.e. the upper and lower 2.5%) of each distribution are indicated by shading; these are also summarised in Table A.3. The solid and dashed grey lines represent the prior and extinct (i.e. those simulations in which all groups became extinct prior to the end of the simulation) density distributions of parameters, respectively.

Figure A.32: NMCI SW FLAT model.

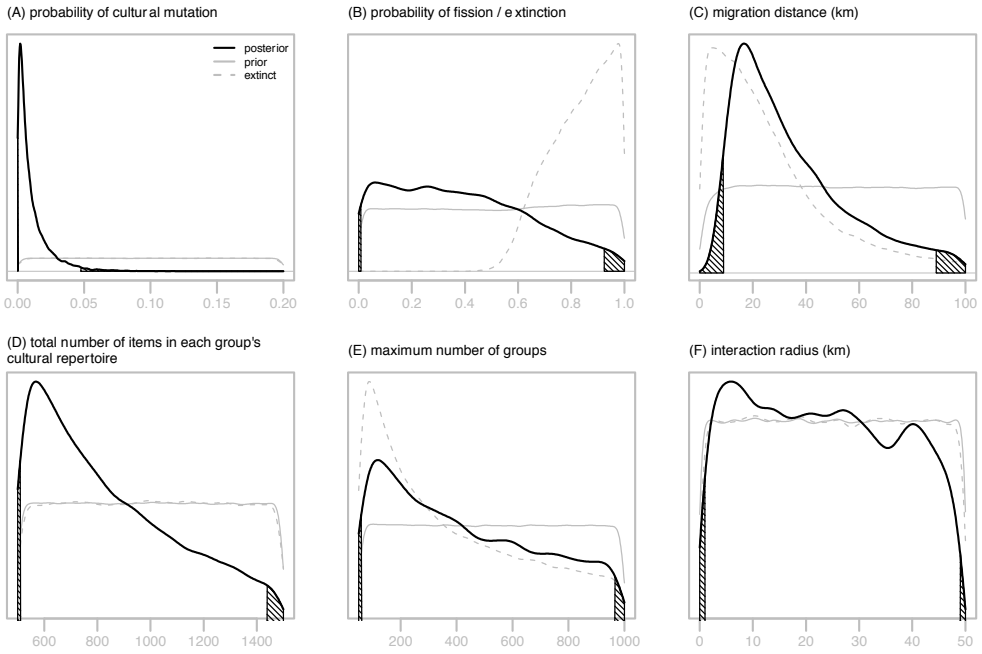


Figure A.33: NMCI SW NPPSingarayer model.

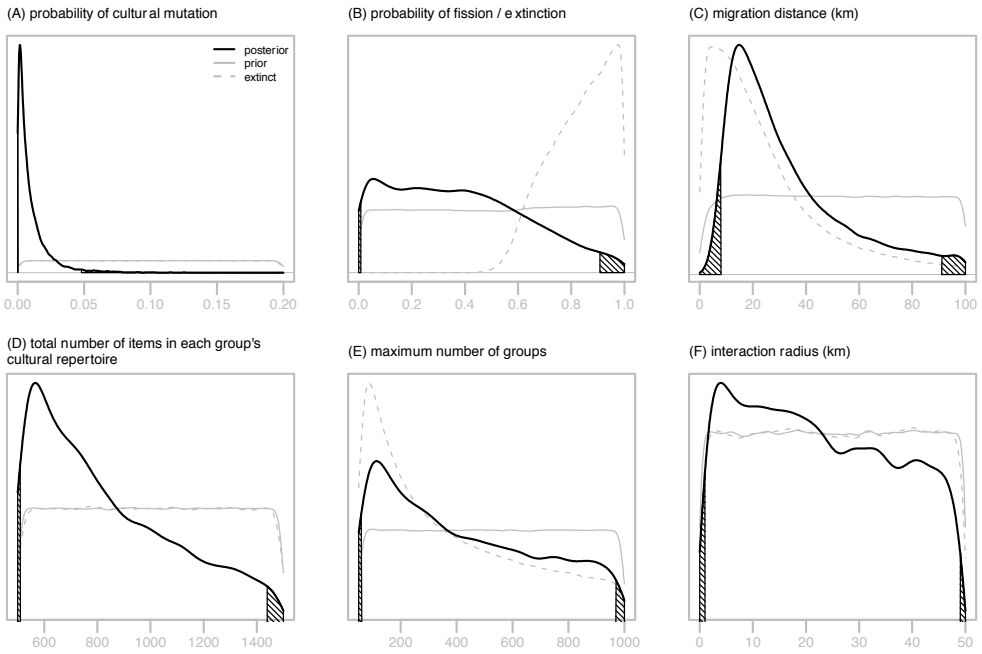
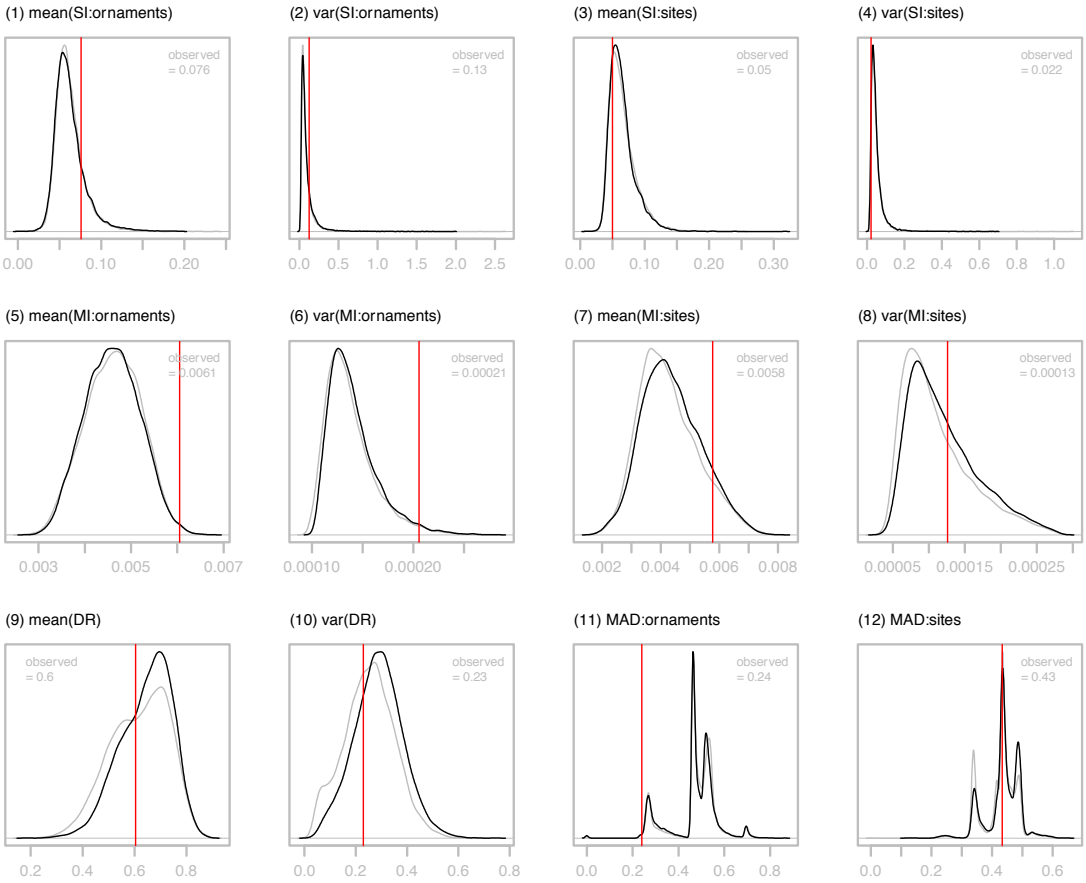


Table A.3: Prior ranges and posterior estimates of parameters for the two versions of interest of the Null Model with Cultural Interactions. For each model, the posterior parameter ranges are calculated on the 12,000 retained simulations (i.e. closest ~1% of 1,132,411 simulations) and expressed by giving the mode, 2.5% and 97.5% quantiles, expressed to 4 decimal places. The letter in the far left column corresponds to the panels in Figure A.32 (NMCI SW FLAT) and Figure A.33 (NMCI SW NPPSingarayer).

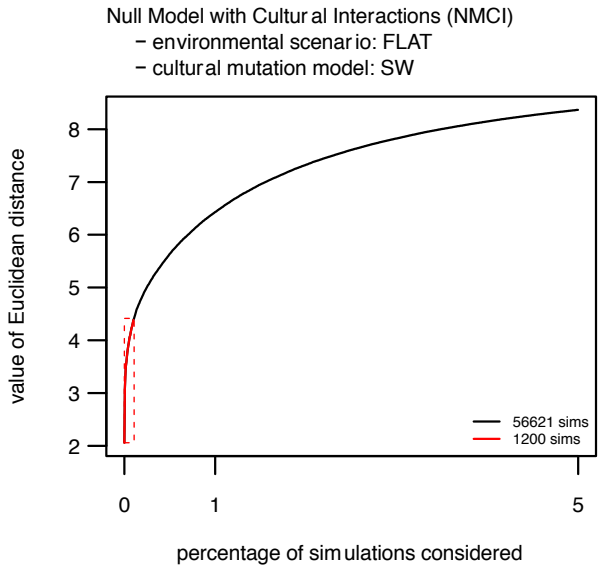
	Parameter	Prior Range		SW FLAT			SW NPPSingarayer		
				Posterior Estimate					
		minimum	maximum	mode	2.5% quantile	97.5% quantile	mode	2.5% quantile	97.5% quantile
(A)	p_{mut}: probability of cultural mutation	0	0.2	0.0020	0.0002	0.0474	0.0016	0.0002	0.0480
(B)	$p_{f/e}$: probability of fission / extinction	0	1	0.0626	0.0091	0.9236	0.0548	0.0080	0.9078
(C)	d_{mig}: migration distance (km)	1	100	16.6341	9	89	14.8728	8	91
(D)	N_{items}: total number of items in each group's cultural repertoire	500	1500	568	511	1438	567	510	1439
(E)	N_{groups}: maximum number of groups	50	1000	119	61	965	115	61	969
(F)	d_{int}: interaction radius (km)	0	50	5.8708	1	49	3.9139	1	49

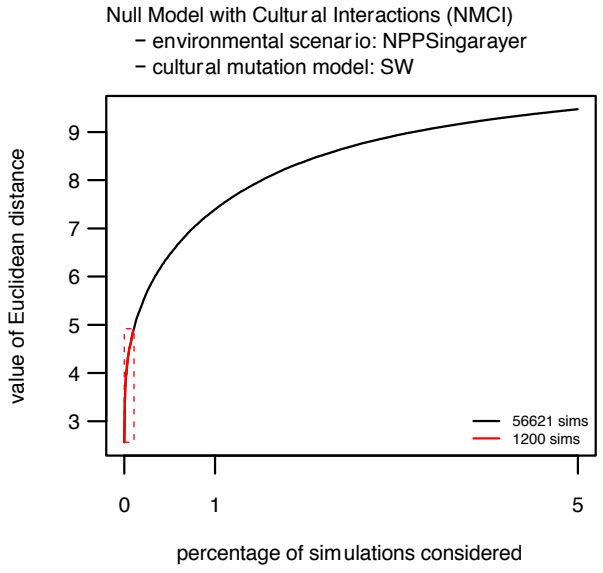
Figure A.34: Distributions of the 12 summary statistic values in the 12,000 retained simulations (i.e. closest ~1% of 1,132,411 simulations) for the NMCI SW FLAT model (black lines) and NMCI SW NPPSingarayer model (grey lines). The title of each panel corresponds to the summary statistic as discussed in section 2.2.1. The red vertical line indicates the target value of each summary statistic (i.e. the value of that statistic calculated from the observed data).



A.4.2 Threshold: 0.1% (i.e. closest 1,200 of 1,132,411 simulations)

Figure A.35: Ranked Euclidean distance values of the best 56,621 simulations (i.e. the closest 5% of 1,132,411 simulations) for each of the two models plotted in black, with the 1,200 retained simulations (i.e. closest ~0.1% of 1,132,411 simulations) used for estimating the posterior parameter distributions of parameters highlighted in red. The two panels correspond to the two models of interest: NMCI SW FLAT (top panel) and NMCI SW NPPSingarayer (bottom panel).





Estimated posterior density distributions of the demographic and evolutionary parameters of interest for the two NMCI models, calculated on the 1,200 retained simulations (i.e. closest ~0.1% of 1,132,411 simulations) for each model. The boundaries of the equal-tailed 95% credible intervals (i.e. the upper and lower 2.5%) of each distribution are indicated by shading; these are also summarised in Table A.4. The solid and dashed grey lines represent the prior and extinct (i.e. those simulations in which all groups became extinct prior to the end of the simulation) density distributions of parameters, respectively.

Figure A.36: NMCI SW FLAT model.

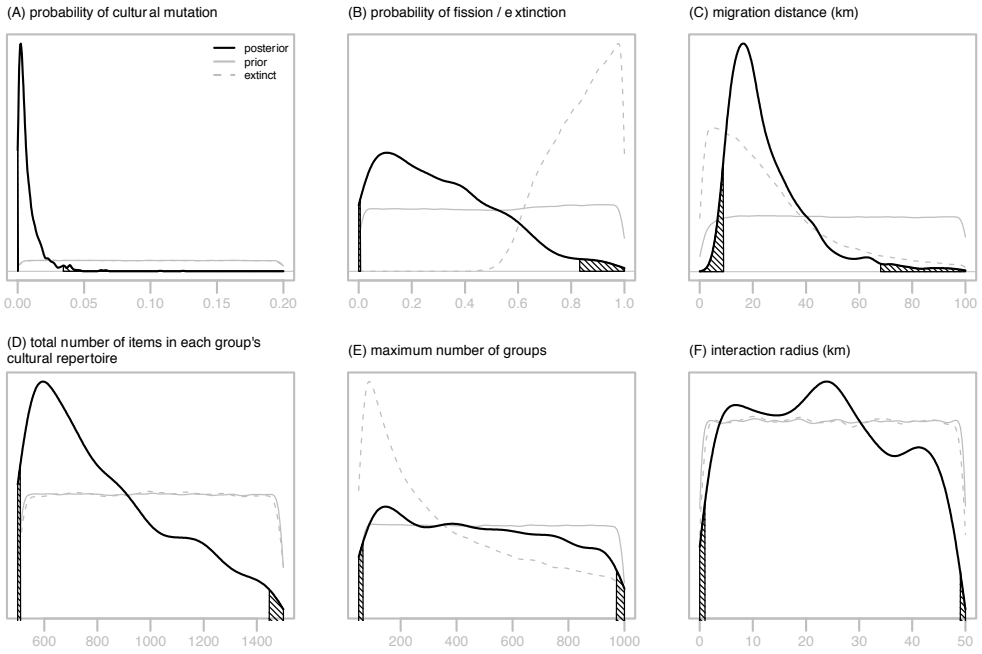


Figure A.37: NMCI SW NPPSingarayer model.

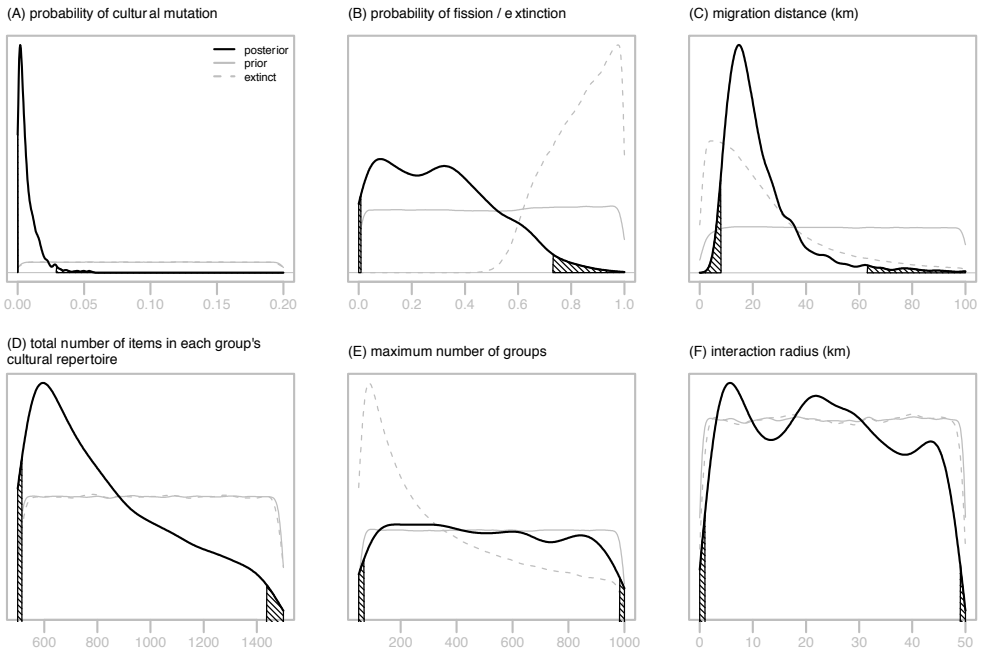
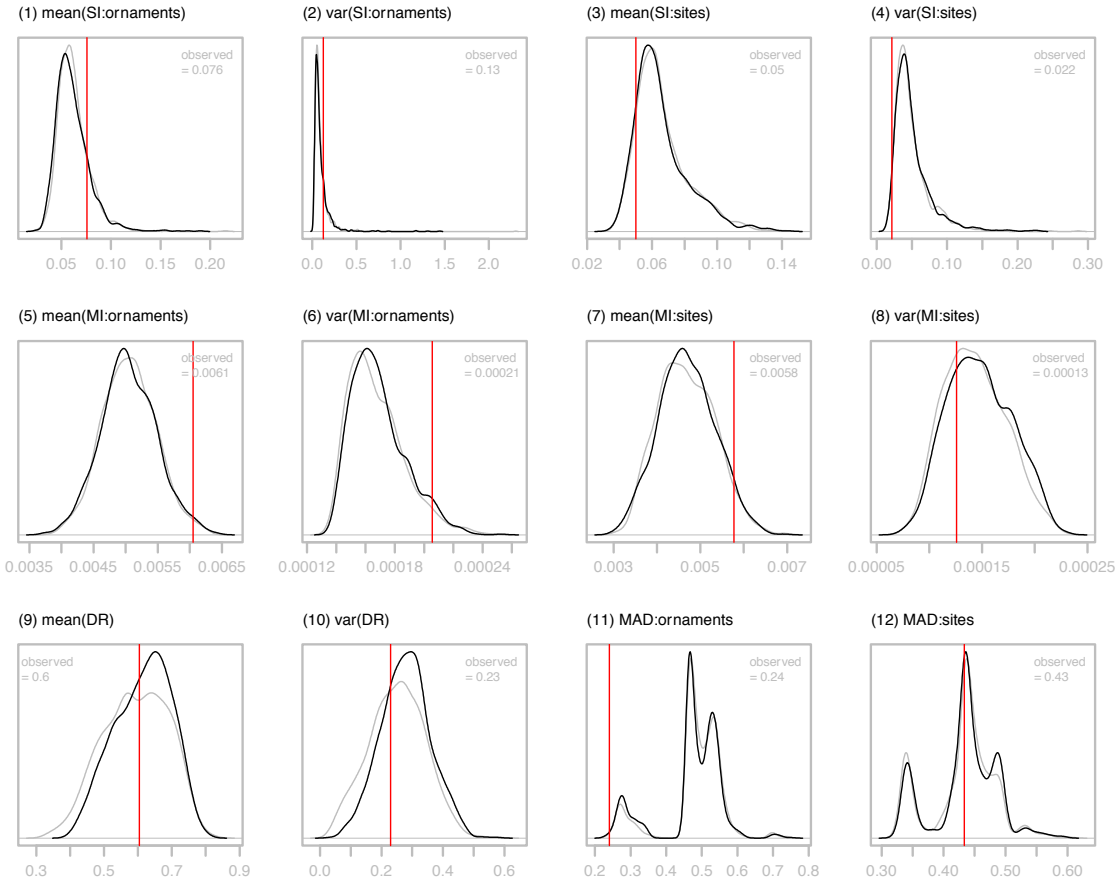


Table A.4: Prior ranges and posterior estimates of parameters for the two versions of interest of the Null Model with Cultural Interactions. For each model, the posterior parameter ranges are calculated on the 1,200 retained simulations (i.e. closest ~0.1% of 1,132,411 simulations) and expressed by giving the mode, 2.5% and 97.5% quantiles, expressed to 4 decimal places. The letter in the far left column corresponds to the panels in Figure A.36 (NMCI SW FLAT) and Figure A.37 (NMCI SW NPPSingarayer).

	Parameter	Prior Range		SW FLAT			SW NPPSingarayer		
				Posterior Estimate					
		minimum	maximum	mode	2.5% quantile	97.5% quantile	mode	2.5% quantile	97.5% quantile
(A)	p_{mut}: probability of cultural mutation	0	0.2	0.0023	0.0003	0.0343	0.0020	0.0002	0.0290
(B)	$p_{f/e}$: probability of fission / extinction	0	1	0.1057	0.0075	0.8313	0.0822	0.0091	0.7309
(C)	d_{mig}: migration distance (km)	1	100	16.4384	9	68	14.6771	8	63
(D)	N_{items}: total number of items in each group's cultural repertoire	500	1500	596	511	1446	596	516	1436
(E)	N_{groups}: maximum number of groups	50	1000	146	66	971	184	70	982
(F)	d_{int}: interaction radius (km)	0	50	23.8748	1	49	5.5.6751	1	49

Figure A.38: Distributions of the 12 summary statistic values in the 1,200 retained simulations (i.e. closest ~0.1% of 1,132,411 simulations) for the NMCI SW FLAT model (black lines) and NMCI SW NPPSingarayer model (grey lines). The title of each panel corresponds to the summary statistic as discussed in section 2.2.1. The red vertical line indicates the target value of each summary statistic (i.e. the value of that statistic calculated from the observed data).



Appendix B: Research Article I

Summary

This section contains the article published as:
Kovacevic M., Shennan S., Vanhaeren M., d'Errico F., Thomas M.G. (2015) "Simulating geographical variation in material culture: Were early modern humans in Europe ethnically structured?", in Aoki K. and Mesoudi A. (eds), Replacement of Neanderthals by Modern Humans, Springer, pp 103-120

Permission to reproduce this article has been granted by *Springer*.

The modelling framework presented in the article was developed and implemented, and analyses on the simulated data performed, by Mirna Kovacevic, with guidance from Mark G. Thomas and Stephen Shennan in a supervisory capacity. Marian Vanhaeren and Francesco d'Errico provided published archaeological data (Vanhaeren and d'Errico 2006).

Simulating Geographical Variation in Material Culture: Were Early Modern Humans in Europe Ethnically Structured?

8

Mirna Kovacevic, Stephen Shennan, Marian Vanhaeren, Francesco d'Errico, and Mark G. Thomas

Abstract

A high degree of structuring is seen in the spatial distribution of symbolic artefact types associated with the Aurignacian culture in Upper Palaeolithic Europe, particularly the degree of sharing of ornament types across archaeological sites. Multivariate analyses of these distributions have been interpreted as indicating ethno-linguistic differentiation (Vanhaeren and d'Errico 2006), although simpler explanations such as isolation-by-distance have not been formally discounted. In this study we have developed a spatiotemporally explicit cultural transmission simulation model that generates expectations of a range of spatial statistics describing the distribution of shared ornament types. We compare these simulated spatial statistics to those observed from archaeological data for Aurignacian Europe—using Approximate Bayesian Computation—in order to test and compare a range of hypotheses concerning group interaction dynamics for the period. Among the set of hypotheses examined, we include ones where material culture does or does not drive group interaction dynamics.

Keywords

Simulation modelling • Culture evolution • Palaeolithic • Demography • Ethnicity

8.1 Introduction

> *"The purpose of models is not to fit the data but to sharpen the questions."* – Samuel Karlin, 11th R A Fisher Memorial Lecture, Royal Society April 1983

Over the last 30 years, there has been a movement from qualitative towards quantitative approaches to the study of

M. Kovacevic (✉)
CoMPLEX (Centre for Mathematics and Physics in the Life Sciences and Experimental Biology), University College London, Physics Building, Gower Street, London WC1E 6BT, UK

Research Department of Genetics, Evolution and Environment, University College London, Darwin Building, Gower Street, London WC1E 6BT, UK
e-mail: m.kovacevic@ucl.ac.uk

S. Shennan
Institute of Archaeology, University College London, 31-34 Gordon Square, London WC1H 0PY, UK

M. Vanhaeren
CNRS (Centre National de la Recherche Scientifique), UMR (Unité Mixte de Recherche) 5199, PACEA (De la Préhistoire à l'Actuel: Culture, Environement et Anthropologie), Université Bordeaux 1, Avenue des Facultés, 33405 Talence, France

F. d'Errico
CNRS (Centre National de la Recherche Scientifique), UMR (Unité Mixte de Recherche) 5199, PACEA (De la Préhistoire à l'Actuel: Culture, Environement et Anthropologie), Université Bordeaux 1, Avenue des Facultés, 33405 Talence, France

Department of Archaeology, History, Cultural Studies and Religion, University of Bergen, Øysteinsgate 3, 7805, 5020 Bergen, Norway

M.G. Thomas
Research Department of Genetics, Evolution and Environment, University College London, Darwin Building, Gower Street, London WC1E 6BT, UK

A. Mesoudi and K. Aoki (eds.) *Learning Strategies and Cultural Evolution during the Palaeolithic*, Replacement of Neanderthals by Modern Humans Series, DOI 10.1007/978-4-431-55363-2_8,

archaeological material culture, and a progression towards modelling approaches to understanding past processes. Archaeologists are now more widely postulating explicit hypotheses to explain the material culture records recovered from archaeological sites, and developing various methods to test these hypotheses. As a result, the fields of archaeology, anthropology and the social sciences in general have become increasingly systematic and multidisciplinary. In archaeology, there has been an increase in the application of computer simulation modelling and statistical techniques to study the relationship between cultural and demographic processes (Clark and Hagemeister 2007; Powell et al. 2009; Costopoulos and Lake 2010; Gerbault et al. 2014) in order to address longstanding archaeological and anthropological questions that are difficult to address through interpretation of archaeological data alone.

The evolution and spread of cultures have been studied using computational modelling methods, with particular focus on processes of cultural innovation and the transmission and accumulation of cultural traits (Neiman 1995; Shennan 2001; Henrich 2004; Powell et al. 2009). The formation of cultural boundaries has also been studied using a group of computational methods labelled agent based modelling (ABM). ABM has been applied throughout the social sciences to investigate how large-scale effects emerge as a result of interactions between agents in the system (Premo 2007; Powell et al. 2009) and for studies of hominin dispersal (Mithen and Reed 2002; Nikitas and Nikita 2005; Hughes et al. 2007). In particular, Robert Axelrod has used ABM methods to investigate the persistence of cultural heterogeneity as a result of interactions between individual agents that are dependent on the extent of cultural similarity between those agents (Axelrod 1997).

Simulation modelling of this kind is a powerful approach that allows the incorporation of stochasticity (variation in demographic and cultural processes arising from random events) into the models. Simulation modelling, and computational modelling in general, also allows researchers to account for sample sizes and the spatial distribution of sample sites, effectively incorporating sampling error and some archaeological bias in inferences on the past (Shennan et al. 2013; Gerbault et al. 2014). The use of modelling in archaeology has resulted in a better understanding of behaviours of agents within the complex systems modelled, as well as helping to refine the questions that are asked and hypotheses that are postulated. With such methods, archaeologists are able to develop robust frameworks that allow a qualitative comparison of alternative modelled scenarios with each other and with observed material culture records, in effect creating virtual experiments to test the effect of varying parameter values on the similarity between simulated and observed material culture data.

In addition to simulation modelling, statistical modelling methods are widely used to describe distributions of, and relationships between, archaeological variables; for example, regression modelling is used to infer correlations between variables of interest. As in many other disciplines, Bayesian methods in archaeology have surged in popularity in recent years. In brief, Bayesian inference is a branch of statistics that uses particular datasets to infer the probability that a proposed hypothesis, or a parameter value of that hypothesis, is true. In contrast to frequentist statistics, where the hypothesis is fixed and variation in outcomes (data) is explored, in Bayesian inference the data becomes fixed and some space of possible explanations (hypotheses) is explored. This means that Bayesian approaches are naturally well suited to archaeological inference since observed data from the past is fixed but only one of a number of possible outcomes of a set of stochastic processes of interest. In Bayesian approaches, various models with set numbers of parameters are proposed, and the posterior probability distributions of these parameters are inferred using information from prior probability distributions of the parameters and information provided by the observed data.

In archaeology, Bayesian methods are primarily associated with dating; for example, to integrate stratigraphic information with radiocarbon date estimates in order to calibrate the probability density distributions (Buck 2001). Other branches of Bayesian methods have not been extensively implemented in archaeological studies. Of particular interest in this paper is a family of Bayesian methods called Approximate Bayesian Computation (ABC) (Tavare et al. 1997; Fu and Li 1997; Beaumont et al. 2002; Bertorelle et al. 2010).

In ABC techniques, a large number of datasets are simulated under a model assuming different, randomly chosen, parameter values from within prior ranges, and appropriate summary statistics are used to measure the extent to which the simulated datasets emulate the observed data. Parameter values under which the model generates datasets closest to the observed data are retained and form a sample of the posterior probability distributions of the parameters. This approach allows the researcher to postulate a number of hypotheses and, provided that they are sufficiently well defined to allow data to be simulated, test which of these hypotheses are more likely given the observed data. An important advantage of ABC over traditional Bayesian approaches is that it is not necessary to formulate an exact function to calculate the probability of the data given some conditions (the likelihood function). The ABC framework and algorithms are further discussed in Appendix 1: Bayesian Inference and Approximate Bayesian Computation (ABC), Appendix 2: Approximate Bayesian Computation (ABC) Algorithm and Appendix 3: Summary Statistics.

In this paper we present a case study in which a spatiotemporally explicit cultural transmission simulation framework has been developed and integrated with observed material culture data (Upper Palaeolithic bead types identified as

personal ornaments), using ABC, in order to aid the interpretation of quantitative data analyses on the observed material culture data (Vanhaeren and d'Errico 2006).

8.2 Case Study: Applying Simulation Modelling and ABC Methods

8.2.1 Introduction

The transition from the Middle Palaeolithic to the Upper Palaeolithic period in Europe occurred as early as approximately 44,000 years ago (Kuhn et al. 2001; Bar-Yosef 2002; Mellars 2005; Higham et al. 2012; Banks et al. 2013). This transition is widely seen as marking the appearance of modern human behaviour in Europe, as evidenced in the Upper Palaeolithic material culture by increased and consistent symbolic activity, and other technological and cultural advances (Powell et al. 2009). These changes in behavioural patterns appear in the archaeological record in the form of abstract and figurative art, the use of personal ornaments, systematically produced microlithic stone tools, bone, ivory and antler artefacts, and increasingly complex hunting technologies. The initial appearance of such items in the European territory dates to the beginnings of the Upper Palaeolithic transition and is thought to coincide with the appearance of AMH in Europe (Kuhn et al. 2001; Zilhão 2007).

The earliest evidence of anatomically modern humans in Europe remains a subject of debate, but is estimated to date to between approximately 45 Ka (Benazzi et al. 2011; Higham et al. 2011) and 40 Ka (Zilhão et al. 2007; Trinkaus and Zilhão 2012). Due to the lack of reliably dated Neanderthal fossils younger than approximately 40 Ka (Pinhasi et al. 2011), archaeological findings dating to 40 Ka or later are assumed to be mostly the result of activities of anatomically modern human populations. Little is known about the migration routes of the first anatomically modern human populations inhabiting Europe at the onset of the Upper Palaeolithic, the extent of biological, cultural and linguistic diversity among them, and the nature and extent of their interactions with the local Neanderthals (but see, for example, Prufer et al. (2014)).

Personal ornaments are considered to be among the first material objects used to communicate social and ethnic identity within and across cultural boundaries (Kuhn et al. 2001). In relation to ethnic identity, personal ornaments can therefore be considered to be the most diagnostic components of material culture surviving in the archaeological record. It has been argued that personal ornaments and beadwork can be used as a proxy for ethno-linguistic identity (Vanhaeren and d'Errico 2006), and that they offer archaeological advantages over other components of the material record for inferring ethno-linguistic structuring, including their exclusively symbolic function, and the frequency and wide assortment in which they occur at archaeological sites associated with the Upper Palaeolithic (Kuhn et al. 2001; Vanhaeren and d'Errico 2006).

In their study, Vanhaeren and d'Errico (2006) considered bead types, identified as personal ornaments, from European Aurignacian sites. Seriation and correspondence analyses of the data identified geographically non-randomly distributed clusters of sites sharing bead types. Seriation analysis is a relative dating method used to chronologically order artefacts recovered from different sites and belonging to the same culture. It is based on the relative chronological order of artefacts and is often applied when absolute dates are not available. Correspondence analysis is related to principal components analysis and is a method used to identify dimensions of variation in categorical data and rank them by the amount of variance explained. The authors argued that the observed variation in spatial distributions was not due to changes over time in personal ornament preference or local availability of raw materials, but rather represented cultural differences among the human groups using Aurignacian technologies. They further argued that the identified trends may have reflected ethno-linguistic diversity among Aurignacian populations.

While this is an interesting interpretation, simpler explanations of these results have not been formally discounted. There are many factors that could cause spatial variation or geographical structuring of material culture, including ethnicity, chronology, local availability of raw materials, environmental influences or simply isolation-by-distance and identity by descent. It is also important to distinguish between spatial variation and ethnic structuring, the latter referring to the ability of individuals, or groups of individuals, to consciously identify with a specific social group *"based on a particular locality or origin"* (Shennan 1989). Considering this definition, it is clear that drawing conclusions about ethnic identity and structure for prehistoric populations is difficult since there are no data in the material record relating to individual's conscious identification; the challenges of invoking ethnic structuring and reconstructing patterns of ethnicity through analysis of material culture data have been discussed by several authors (Shennan 1989, 2002; Jones 1997). With this in mind, invoking ethnic structure for the Upper Palaeolithic in Europe is a challenging task given the paucity of material culture and other data for the period.

However, ethnic identity and structuring are universals in the modern world and are therefore frequently assumed for peoples in the past. Identifying the earliest appearance of ethnicity is an issue of general importance for the history of human evolution that has implications for the emergence of languages, and may inform on the evolutionary dynamics of human populations, as well as the role of identity construction in people today.

The current study therefore aims to test whether the distribution of artefact types reported by Vanhaeren and d'Errico (2006) can be explained by a model of cultural identity-by-descent with modification and isolation-by-distance, or whether it is necessary to invoke cultural group interaction processes that would be expected if material culture was symbolically marking ethnic group identity. For example, an interaction between two culturally similar populations may result in sharing of cultural traits between the two, causing them to become overall more culturally similar, while an interaction between two culturally different populations might result in the two undergoing conflict, dependent on the extent of the cultural difference between them, and possibly the imposition of one culture on another. An analogous distinction is that between the existence of inter-group differences arising through cultural mutation and drift (the null model), of which actors are not consciously aware, versus that null model plus the intentional adherence to behavioral norms that imply identity and actively shape interaction processes, and through that, the spread, loss and mixing of culturally inherited traits.

In this study, spatiotemporally explicit cultural transmission simulation models that generate simulated material culture data under each of the scenarios described above have been developed and explored through simulation. The archaeological dataset published by Vanhaeren and d'Errico (2006) is used to assess the validity of each model. The underlying principle here is that conditions under which the simulated data is very similar to the observed archaeological data—as reflected in a range of spatial statistics describing the distribution of artefact types—are more likely to be true than conditions under which the simulated data is unlike the observed data. This assessment of the goodness-of-fit between the simulated and observed data is quantified using ABC.

8.2.2 Simulation Modelling

Each simulation is initialised at the onset of the Aurignacian period, approximately 42 Ka, and simulated forward in time to the end of the Aurignacian period, approximately 29 Ka (Higham et al. 2012). Each simulation spans a total of 13,000 years, or 520 generations assuming a 25 year generation time (Tremblay and Vezina 2000; Thomas et al. 2006). Since this may be an overestimate of the length of the Aurignacian period (Zilhão and Pettitt 2006), data is also collected when each simulation reaches 10,000 years, or 400 generations, though these results are not presented here. Each simulation includes a 1,000 year, or 40 generation, burn-in period at the start of the simulation during which no simulated data is collected, in order to allow for possible inaccuracies in initial locations of simulated groups.

8.2.2.1 Simulation World

The geographic region considered in this study is the range of latitudes and longitudes corresponding to the European territory. The longitude, ϕ, ranges from $-11°$ to $30°$, which, relative to modern day country boundaries, is approximately the area from the western Irish boundary to the western Russian boundary at the Urals. The latitude, λ, ranges from $35°$ to $60°$, which is approximately the area from the northern boundary of Africa to the northern boundary of Scotland. Although it would be possible to incorporate changes in sea levels through time by using available bathymetry data, dramatic geostatic rebound for northern latitudes makes it difficult to accurately estimate coastlines for northern Europe. For this reason, modern coastlines are currently used in simulations.

Within the defined region, each geographic location is assigned a local carrying capacity. The carrying capacity of a location determines the habitability, and therefore potential population density, of that location; a zero carrying capacity corresponds to an uninhabitable region, for example sea or ice covered land. In order to estimate these local carrying capacities for geographic locations in the modelled domain, two distinct environmental scenarios have been considered; each simulation is conditioned on only one of these two environmental scenarios.

The first is a simple scenario in which Europe is assumed to be a flat space. This corresponds to a distribution with all locations within the modelled domain having equal relative carrying capacity values.

In the second of the environmental scenarios, instead of treating Europe as a flat space, we have taken information on estimated population densities during the Aurignacian from Bocquet-Appel et al. (2005) to inform on carrying capacities for the modelled domain, shown in Fig. 8.1. Bocquet-Appel et al. used databases of archaeological sites corresponding to the Upper Palaeolithic period, together with simulated climatic variables and ethnography of hunter-gatherers, to estimate the distribution of hunter-gatherer populations in Upper Palaeolithic Europe.

The original estimate of population density, shown in Fig. 8.1 for the Aurignacian period, was not made available, so the distribution used in this study is approximated based on the original figure. Since we are concerned with distributions rather than exact numbers estimated in the original study (Bocquet-Appel et al. 2005), this estimate is normalised to give relative distributions. The normalised distribution is used in simulations as the relative carrying capacity value for each location.

In both scenarios, the potential, or target, population density for each location is calculated as the product of the relative carrying capacity value at that location and the G_{max} parameter, which specifies the total maximum number of groups that the modelled domain can sustain. We treat this as an unknown parameter and explore a range of values.

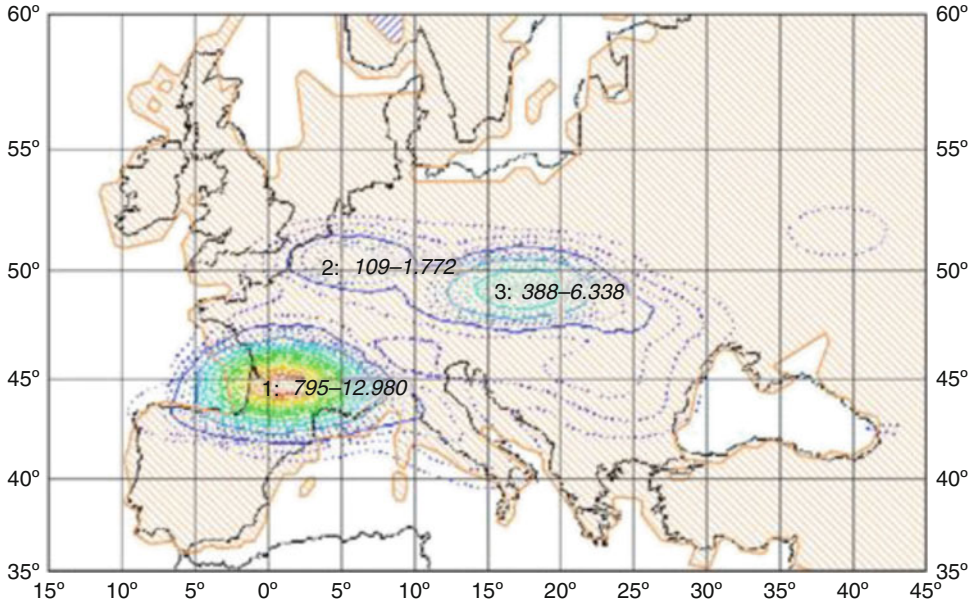


Fig. 8.1 Estimate of the regional distribution of the metapopulation of hunter-gatherers during the Aurignacian period of the Upper Palaeolithic in Europe, superimposed on the IOS3 project maps. The boundaries (*in black*) of the accretion zones, with the corresponding numbers, account for roughly 90 % of the distribution of the local population (Image and edited caption from (Bocquet-Appel et al. 2005)

8.2.2.2 Demographic Processes

Each simulation is initialised with a fixed number of groups, G_0, placed in randomly chosen habitable locations in the modelled domain; all attributes and processes are defined at the level of the group, rather than individuals in that group, and groups are assumed to be the same size. Groups migrate locally and undergo fission/extinction processes. These demographic processes are analogous to an isolation-by-distance model in population genetics (Wright 1943; Slatkin 1993).

Migratory Processes

At each generation groups are subjected to migratory processes modelled as parameterised Gaussian random walks. The distance each group traverses in a migration process is picked from a normal distribution with mean μ_{mig} and standard deviation $\sigma_{mig} = d_{mig}$. Positive and negative values picked from the distribution correspond to movement in opposite directions, namely East and West and North and South, respectively. The mean of the distribution is therefore set to $\mu_{mig} = 0$ to ensure that movement in opposite directions is equally likely. Parameter d_{mig} corresponds to the standard deviation, or width, of the normal distribution and specifies the range of values that the migration distance is most likely to take in each of the East-West and North-South directions. We treat d_{mig} as an unknown parameter and explore a range of values.

The distance travelled by each group at each generation in the East-West and North-South directions is picked independently from the above-described normal distribution. The distance, d, and direction, θ, that define each group's movement are given by:

$$d = \sqrt{\left(\Delta x\right)^2 + (\Delta y)^2}, \text{ and} \tag{8.1}$$

$$\theta = \arctan 2\left(\Delta x,\ \Delta y\right), \tag{8.2}$$

where *arctan2* corresponds to a variant of the *arctan* function that takes into account the sign of both vectors in question and distinguishes diametrically opposite directions, therefore specifying unique angle values in the range $(0, 2\pi)$.

The new proposed position of each group is then calculated based on the group's current location, the distance, d, and the direction, θ, of movement. If the longitude and latitude of the group's current positions are $\phi_{current}$ and $\lambda_{current}$, respectively, and the same for the group's new locations are ϕ_{new} and λ_{new}, respectively, then:

$$\lambda_{new} = \sin^{-1}\left(\sin\left(\lambda_{current}\right) \cdot \cos\left(\frac{d}{R}\right) + \cos\left(\lambda_{current}\right) \cdot \sin\left(\frac{d}{R}\right) \cdot \theta\right), \tag{8.3}$$

$$\phi_{new} = \phi_{current} + \arctan 2\left(\cos\left(\frac{d}{R}\right) + \cos\left(\lambda_{current}\right) \cdot \sin\left(\frac{d}{R}\right) \cdot \theta,\ \cos\left(\frac{d}{R}\right) - \sin\left(\lambda_{current}\right) \cdot \sin\left(\lambda_{new}\right)\right). \tag{8.4}$$

These formulae are introduced to allow for the curvature of the Earth when calculating the new group positions. Although the curvature of the Earth has little effect in the current framework, as the migration distance d is small

relative to the radius of the Earth (denoted R and assumed to be constant at 6,371 km) using these formulae ensures that the model can be applied accurately with arbitrarily large migration distances.

Fission/Extinction Processes and KDE

In addition to the migratory process undergone at each generation, each group also undergoes a fission/extinction process with parameterised probability. The probability that a group undergoes a fission/extinction process is given by the probability of fission/extinction parameter, $p_{f/e}$; we treat this parameter as an unknown and explore a range of values. The type of process that a selected group undergoes is determined by the difference between target and current local population density at the group's location (i.e. fission or extinction). The difference between target and current local population density is an indicator of potential for growth; a positive value indicates that the location is below carrying capacity (i.e. the target local population density is greater than the current local population density—the location is under-populated and so there is potential for growth) and therefore results in a fission event, while a negative value indicates that the location is above carrying capacity (i.e. the target local population density is smaller than the current local population density—the location is over-populated and so there is no potential for growth) and therefore results in an extinction event. The population density at the current generation is estimated from the group locations using kernel density estimation (Wand and Jones 1995).

An extinction event results in the group being deleted from the simulation, while a fission event results in a replication such that two groups, the parent and offspring, are present in the next generation. The offspring group retains the cultural traits of the parent group (i.e. the offspring group is an exact replica of the parent group, except for any mutation events), analogous to identity-by-descent in population genetics. In subsequent generations, the parent and offspring groups migrate and undergo fission/extinction processes independently, and their respective cultures also evolve independently.

8.2.2.3 Cultural Processes: Modelling Ethnic Diversity

Axelrod's Model of Cultural Dissemination

The models developed in this study simulate innovation in culture and so require the concept of culture to be mathematically defined. For this purpose we have used an adapted version of Axelrod's definition (Axelrod 1997) in which the culture of an agent (an individual or a group of individuals) is defined to be a set of attributes that are subject to social influence. In Axelrod's definition, the culture of an agent consists of some number of these attributes, referred to as cultural *features*, and each can assume one of a predefined number of values, referred to as *traits*, thus, each agent is monomorphic for each cultural feature. In this definition, the culture of an agent is then described as a list of digits, with the position of a digit corresponding to the feature and the value of a digit specifying the current trait for that feature. In Axelrod's definition, the trait—or value that a feature takes—is assigned at the start of the simulation and is only influenced by social interactions (i.e. it does not undergo any mutation processes).

In Axelrod's formulation, social interactions are constrained to occur only between agents that are immediate neighbours. The simulations occur on a square lattice with agents arrayed at discrete points over the lattice. Most agents therefore have four immediate neighbours, with those on the edge of the lattice having three and those in the corners having two immediate neighbours. Also in Axelrod's model, the probability of an interaction between two agents is proportional to the cultural similarity between them. This similarity is quantified as the proportion of their features that have the same trait. The interaction then consists of an agent, and an immediate neighbour to that agent, being chosen at random. A single feature on which the chosen agent's culture and the neighbour's culture differ is selected at random, and the value of this feature (trait) in the chosen agent's culture is set to the value of the same feature in the neighbour's culture.

This formulation is a good basis; however, it is very limited in diversity of cultural features and traits and is inadequate to capture the high dimensionality of the observed data used in the current study (Vanhaeren and d'Errico 2006). In addition, over long chronological periods such as those simulated here, it is necessary to consider the effect of cultural mutation and drift processes. This definition must therefore be modified so that it can be applied to the current problem.

Observed and Simulated Datasets

The observed dataset (Vanhaeren and d'Errico 2006) consists of 157 distinct bead types recorded at 98 Aurignacian sites in Europe and the Near East, with records specifying presence/absence of distinct bead types in sites only. These distinct bead types are divided between 11 features according to different raw materials, with *"62 representing ornaments made of shells, 31 of teeth, 30 of ivory, 11 of stone, 11 of bone, 7 of deer antler, and one each of belemnite, nummulite, ammonite, sea urchin and amber"* (Vanhaeren and d'Errico 2006).

In the models developed in this study, we have adapted Axelrod's definition of culture described above so that each agent, in our case a group, is polymorphic for each cultural feature. To allow for this, each group carries a parameterised number of items, or beads, in its cultural repertoire, specified by the N_{items} parameter, treated as an unknown and chosen at the onset of each simulation from a pre-defined range of

values. These items are then divided between the 11 features probabilistically (using a multinomial function), such that the probability of an item being assigned to a particular feature is proportional to the ratio of unique items observed in that feature and the total number of unique items observed (39.5 % shells, 19.7 % teeth, 19.1 % ivory, 7 % stone, 7 % bone, 4.5 % deer antler, and 0.6 % each for belemnite, nummulite, ammonite, sea urchin and amber). Within each feature, each item can then take one of a number of unique possible values, corresponding to the number of distinct bead types for that feature in the observed ornament data (Vanhaeren and d'Errico 2006).

Mutation and Drift

The culture of each group undergoes mutation and drift processes at each generation, such that the culture of each group at the next generation will be the combined result of mutation and drift processes acting on the culture of that group at the current generation. In addition to testing various environments as described above, two different models of cultural variation have been considered; in each simulation data is simulated under only one of these two cultural variation models.

In the first, mutation is modelled according to the bounded stepwise model often used to model mutations at microsatellite loci in population genetics (Kimura and Ohta 1978; Valdes et al. 1993), and occurs at each generation for each item in each group's culture with probability proportional to the p_{mut} parameter. We treat this parameter, which specifies the probability of mutation, as an unknown and explore a range of possible values. Under this stepwise mutation model, a cultural trait in a particular feature at the current generation is constrained to mutate to one of the cultural traits on either side of it, within that feature, at the next generation—mutation therefore changes the frequency with which each trait occurs in the next generation. In this case, we assume that cultural traits are ordered in such a way that adjacent traits are more similar than traits that are further apart in the sequence. Since cultural traits considered in this study are discrete and fixed (i.e. one trait cannot morph into another trait), this stepwise mutation model corresponds to a group being more likely to add an item to its cultural repertoire that is morphologically similar to one that is already present in its cultural repertoire than one that is very different. Similarly to population genetics, cultural mutation has the effect of increasing diversity.

In the second, mutation is discrete within the bounds of each feature. Similarly to the stepwise mutation model, in this bounded discrete model mutation occurs at each generation for each item in each group's culture with probability proportional to the p_{mut} parameter. This parameter again specifies the probability of mutation; it is treated as an unknown and a range of possible values are explored. Under this mutation model, however, a cultural trait in a particular feature at the current generation is permitted to mutate to any of the other cultural traits within that feature with equal probability at the next generation. The mutation process again changes the frequency with which each trait occurs in the next generation and has the effect of increasing diversity.

Drift has the opposite effect and decreases the amount of diversity in each group's culture. It is modelled based on genetic drift, where allele frequencies change as a result of random differences in reproduction; in finite populations drift corresponds to the intergeneration sampling error (see, for example, Tishkoff and Verrelli (2003)). The drift process is modelled by using a multinomial function to sample the traits of each cultural feature independently. This implementation takes into account frequencies of cultural traits in the current generation, such that, for a particular group, cultural traits that are at higher frequencies in the group's culture at the current generation are more likely to be present in the group's culture at the next generation.

Depositing Cultures in Sites

Locations of sites in the model are defined to correspond to the locations of the archaeological sites in the observed data (Vanhaeren and d'Errico 2006). A group will deposit its culture at a site when within a specified geographic distance of that site. This catchment distance is initially set to be equal for all sites, with the further constraint that if two groups are within the catchment distance then the group closest to the site will be the one to deposit its culture there.

The distance measure used to calculate the distance between group locations and archaeological sites is the geodesic distance, which is the aerial path between two points, also called the as-the-crow-flies, great-circle or orthodromic distance. To account for curvature of the Earth, geographic distances are calculated using the Haversine Formula (Sinnott 1984). This calculates the great-circle distance between two points on a sphere given their respective longitudes and latitudes. If the longitude and latitude of the points are ϕ_1 and λ_1 for point one and ϕ_2 and λ_2 for point two, respectively, and:

$$\Delta\phi = \frac{\phi_1 - \phi_2}{2}, \tag{8.5}$$

$$\Delta\lambda = \frac{\lambda_1 - \lambda_2}{2}, \tag{8.6}$$

then the distance, D, between the two points is calculated as:

$$D = 2R \sin^{-1}\left(\left(\sin^2(\Delta\phi) + \cos(\phi_1)\cdot\cos(\phi_2)\cdot\sin^2(\Delta\lambda)\right)^{\frac{1}{2}}\right), \tag{8.7}$$

where R is the radius of the Earth, assumed to be constant at 6,371 km.

Each site is assigned an item capacity, which corresponds to the number of items recovered from that site as reported in the original study (Vanhaeren and d'Errico 2006). When a group comes within the specified distance for a particular site, a number of unique items, equal to the item capacity for that site, are selected probabilistically (using a multinomial function so that trait frequencies are taken into account) from the group's entire culture to be deposited at the site—this corresponds to one copy of each cultural trait that is selected being deposited at the site. The original dataset contains presence/absence records of distinct bead types only, so specifying the item capacity and 'uniqueness' of items deposited at each site in simulations should theoretically minimise archaeological and sampling bias. A group arriving at a site at which a deposit has previously been made will deposit its culture at that site and overwrite the existing deposit only if it comes within closer proximity to that site than the last group that deposited its culture there.

Simulated material culture data deposited at the locations of the archaeological sites in the observed data (Vanhaeren and d'Errico 2006) are collected at the end of each simulation, which corresponds to the end of the Aurignacian period.

Cultural Interactions

A pair of groups will interact if they are within a parameterised geographical distance, d_{int}, of each other; we treat this parameter as an unknown and explore a range of values. As above, the distance between groups is calculated as the geodesic distance and the Haversine Formula (Sinnott 1984) is used to account for curvature of the Earth. Two cultural interaction processes are modelled in this framework—conflict and sharing—and a pair of interacting groups will undergo one of these two processes.

The outcome of a conflict process is the replacement of the culture of one group by that of the other. To model this, we assign one in each pair of interacting groups as a winning, and one as a losing group, and replace the culture of the losing group entirely by that of the winning group. The decision on which is assigned to be the winning group, and which the losing group, is made at random due to the assumption that the aspects of material culture we are considering (personal ornamentation) do not have an effect on, and are not a proxy for, group fitness. Additionally, since groups are modelled such that they are the same size, group size cannot be used as a proxy for group fitness. The conflict interaction process is analogous to a group imposing its culture on a group that they have defeated, or, alternatively, assimilating the defeated group into their own, followed by a fission process.

The other interaction process considered is sharing of cultures between interacting groups. Sharing is modelled by permutation, whereby the cultures of the two interacting groups are pooled, permuted and then divided between the two. This is analogous to culturally similar groups swapping cultural traits.

8.2.2.4 Null and Culture-Dependent Interaction Models

The Null Model and Culture-Dependent Interaction Model are both models of cultural identity-by-descent with modification and isolation-by-distance, and are made up of the demographic and cultural processes described above.

The difference between the two models lies in the method of deciding which type of interaction will occur between two interacting groups. In the Null Model, the type of interaction is decided at random; groups are equally as likely to share material culture as they are to undergo conflict. The Null Model is therefore a scenario in which group interactions are independent of similarities or differences in groups' ornamental material cultures. Conversely, in the Culture-Dependent Interaction Model, the type of interaction is decided probabilistically and depends on the extent of cultural similarity between the two interacting groups; groups that are relatively culturally similar are more likely to share cultures while those that are relatively culturally different are more likely to undergo conflict. The main aim of this study is to test which of these two models best explains the observed spatial distribution of ornament types in the archaeological record; the latter is intended to represent the effects of ethnic structuring on the spatial distribution of material culture.

Measures of Cultural Similarity

The extent of cultural similarity between a pair of interacting groups is quantified differently depending on which of the models of cultural variation described above is considered. In simulations that follow the stepwise mutation model, the extent of cultural similarity is quantified using a measure akin to the $(\delta\mu)^2$ measure used to quantify the genetic similarity between populations using microsatellite data (Goldstein et al. 1995a, b). We define this cultural $(\delta\mu)^2$ measure as:

$$(\delta\mu)^2 = \sum_i \sum_j (i-j)^2 x_i y_j - \frac{1}{2}\left[\sum_i \sum_j (i-j)^2 x_i x_j + \sum_i \sum_j (i-j)^2 y_i y_j\right], \tag{8.8}$$

where x_i and y_j are frequencies of traits i and j in (interacting) groups x and y respectively. This measure therefore quantifies the cultural similarity between the two interacting groups by taking into account the frequencies with which all traits occur in their respective cultural repertoires—this measure does not discriminate between differences in cultural features. The

calculated value of $(\delta\mu)^2$ is normalised by the maximum $(\delta\mu)^2$ recorded up to that generation of that simulation, giving a measure of cultural similarity that is relative to the maximum measured cultural similarity.

In simulations that follow the non-stepwise mutation model, the extent of cultural similarity between a pair of interacting groups is quantified using a measure akin to the F_{ST} measure used to quantify the genetic similarity between populations using allele frequency data (Wright 1978; Cavalli-Sforza et al. 1996). We define the cultural F_{ST} measure as:

$$F_{ST} = \frac{H_T - \overline{H_S}}{H_T}. \tag{8.9}$$

In this definition, H_T is the amount of variation in all traits in the whole population (considered to be the two interacting groups) and is defined as:

$$H_T = 1 - \sum_i \overline{p_i}^2, \tag{8.10}$$

where $\overline{p_i}$ is the average frequency of trait i calculated over the two interacting groups. H_S is the amount of variation between traits within each group (calculated separately for each of the two interacting groups); $\overline{H_S}$ is the average of H_S calculated over the two interacting groups. H_S is defined as:

$$H_S = 1 - \sum_i {p_i}^2, \tag{8.11}$$

where p_i is the frequency of trait i. Similarly to the $(\delta\mu)^2$ measure discussed above, the F_{ST} measure takes into account the frequencies with which all traits occur in the cultural repertoires of the two interacting groups—as with the $(\delta\mu)^2$ measure, this measure does not discriminate between differences in cultural features. It is therefore an estimate of the proportion of the total variation in a set of traits that is the result of between-group differences (Bell et al. 2009). Similarly again to $(\delta\mu)^2$, the calculated value of F_{ST} is normalised by the maximum F_{ST} recorded up to that generation of that simulation, giving a measure of cultural similarity that is relative to the maximum measured cultural similarity.

The relative values of $(\delta\mu)^2$ (bounded stepwise mutation model) and F_{ST} (bounded discrete mutation model) can take values between 0 and 1 and are treated as probabilities to decide which of the interaction processes described above occurs between the two interacting groups; a value of 0 indicates that the two groups have identical cultural repertoires and are therefore more likely to share cultures, while a value of 1 indicates complete cultural difference and indicates that the two groups are more likely to undergo conflict.

8.2.2.5 Models, Model Parameters and Prior Ranges

Given that one of two environmental scenarios and one of two models of cultural variation are considered for each simulation in both the Null Model and the Culture-Dependent Interaction Model, data is simulated under eight distinct scenarios. These are summarised in Table 8.1. The acronym and text colour associated with each model correspond to those used in Fig. 8.2 for that model.

In total there are six parameters that govern the processes considered in the Null and Culture-Dependent Interaction Models. Both models have 4 key processes: migration, fission/extinction, cultural mutation and cultural interaction, governed by 4 parameters: d_{mig}, $p_{f/e}$, p_{mut} and d_{int}, respectively. In addition to these, there are two further parameters in both models, namely the maximum number of groups, G_{max}, and the number of items in each group's culture, N_{items}.

There is little information in the archaeological record relating to the precise values that these parameters may take. Each parameter is therefore constrained to a uniform prior range, with the value of each parameter in each simulation randomly assigned from this uniform prior. Prior ranges for each parameter are listed in Table 8.2.

8.2.3 Analysis

Once a large number of simulations have been performed under the models described above, the objective of the data analysis is to quantify the extent of similarity between observed and simulated material culture data. To do this, ABC techniques are used to compare the differences in goodness-of-fit between the observed data (Vanhaeren and d'Errico 2006) and data simulated by different proposed models. To be able to compare the observed and simulated datasets, robust statistics that sufficiently describe the full properties of the data, referred to as summary statistics, are used. Summary statistics used in this study are discussed in detail

Table 8.1 Summary of combinations of environmental scenarios and cultural variation models under which data is simulated in both the Null Model and Culture-Dependent Interaction (CDI) Model

Environmental scenario / Cultural variation model	Flat space (FLAT)	Bocquet-Appel et al. (2005) distribution (B-A)
Bounded stepwise mutation model (SW)	Null model - - - - - CDI model ———	Null model - - - - - CDI model ———
Bounded discrete mutation model (DIS)	Null model - - - - - CDI model ———	Null model - - - - - CDI model ———

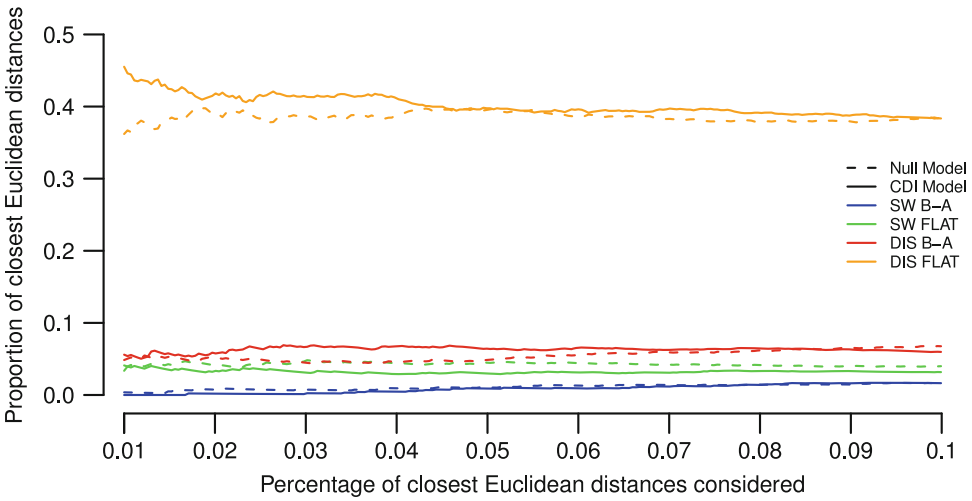


Fig. 8.2 Relative marginal likelihood estimates (y-axis) of each Null Model (*dashed lines*) and Culture-Dependent Interaction Model (*solid lines*) for each percentage (x-axis) of closest simulations, taking into consideration 2,680,000 simulations (335,000 simulations for each Null Model and Culture-Dependent Interaction Model)

Table 8.2 Model parameters and their prior ranges, for both the Null Model and the Culture-Dependent Interaction Model

		Prior range	
Parameter		Null Model	Culture-Dependent Interaction Model
Migration distance (km)	d_{mig}	[0,100]	[0,100]
Probability of fission/extinction	$p_{f/e}$	[0, 1]	[0, 1]
Probability of cultural mutation	p_{mut}	[0, 0.2]	[0, 0.2]
Number of items	N_{items}	[500, 1,500]	[500, 1,500]
Maximum number of groups	G_{max}	[50, 1,000]	[50, 1,000]
Interaction radius (km)	d_{int}	[0, 50]	[0, 50]

in Appendix 3: Summary Statistics. By comparing summary statistics calculated for each simulated dataset to those for the observed data, this method allows us to accept those simulations with summary statistics sufficiently close to the target summary statistics—these are the best simulations, that is, those generating data most similar to the observed data.

Another useful feature of this approach is the ability to formally compare the performance of different models using Bayes Factors (Kass and Raftery 1995). In short, a Bayes Factor is a summary of the evidence provided by the data in favour of one model over another; this is further discussed in Appendix 4: Bayes Factors for Model Comparison. What we are estimating in this study are the relative marginal likelihoods of each proposed model given the data. More explicitly, given models M_1 and M_2 that we want to compare, their respective relative marginal likelihoods l_1 and l_2 are defined as:

$$l_1 = \frac{N_1}{N}, \text{ and} \tag{8.12}$$

$$l_2 = \frac{N_2}{N}, \tag{8.13}$$

where N_1 and N_2 are the number of simulation that come from models M_1 and M_2, respectively, and N $(=N_1 + N_2)$ is the total number of simulations considered; the relative marginal likelihood of each model is defined to be the proportion of total number of simulations considered that come from that model. This is therefore a measure of which model explains the observed data better, given that N simulations are considered.

This form of model comparison is independent of the number of parameters for each model, and instead estimates the likelihood of the model considering all possible parameter values. In cases where models with different numbers of parameters are compared, this method automatically and correctly penalises model complexity; for models with a large number of parameters there is a larger parameter space to explore and so it is more difficult to find those parameter sets that generate data similar to the observed data. Models with more parameters are therefore penalised for the increased complexity compared to simpler models, resulting in a comparison weighted by model complexity. Such an approach prevents us from overfitting—from invoking parameters to explain aspects of the data that are in fact due to randomness. However, in this particular study the number of parameters is equal in all models.

8.2.4 Results

Results shown are from 335,000 simulations for each of the Null and Culture-Dependent Interaction Models. For the relative marginal likelihood estimation, results for the

eight models are considered together; a total of 2,680,000 simulations are therefore taken into account. In this analysis, we estimate the relative marginal likelihood of each model, taking into account the extent of similarity between simulated and observed data (Vanhaeren and d'Errico 2006).

Figure 8.2 is a plot of the estimated relative marginal likelihood of each version of the Null Model (dashed lines) and Culture-Dependent Interaction Models (solid lines) at different thresholds. It shows what proportion (y-axis) of the best simulations—those generating data most similar to the observed data—are coming from each model for each percentage (x-axis) of closest simulations. The colours refer to the combination of the environmental scenario and the model of cultural variation considered, as detailed in Table 8.1. Since the plot depicts proportions, for any particular percentage of closest simulations (i.e. for any particular value on the x-axis), the sum of the proportions of the closest Euclidean distances coming from each model (i.e. the sum of the values on the y-axis) will always be 1. The relative marginal likelihood estimates of each version of Null Model and Culture-Dependent Interaction Model for 0.1 % of closest simulations (i.e. Fig. 8.2, x = 0.1) are also detailed in Table 8.3.

Figure 8.2 indicates that, for all scenarios modelled (scenario here referring to a pairwise combination of an environmental scenario and a model of cultural variation as explained in Models, Model Parameters and Prior Ranges), there is little difference in how well the Null Model and Culture-Dependent Interaction Model perform. The best fits of simulated to observed data are generated by data simulated under the scenario that combines the bounded discrete mutation model and the environmental scenario in which Europe is assumed to have a flat distribution of carrying capacities (represented by orange lines in Fig. 8.2), with approximately 38.4 % of the best 0.1 % of simulations coming from each the Null Model (dashed orange line) and Culture-Dependent Interaction Model (solid orange line).

Table 8.3 Relative marginal likelihood estimate of each Null Model and Culture-Dependent Interaction (CDI) Model for 0.1 % of closest simulations

Environmental scenario / Cultural variation model	Flat space (FLAT)	Bocquet-Appel et al. (2005) distribution (B-A)
Bounded stepwise mutation model (SW)	Null model: 4.0% CDI model: 3.2%	Null model: 1.6% CDI model: 1.6%
Bounded discrete mutation model (DIS)	Null model: 38.4% CDI model: 38.4%	Null model: 6.8% CDI model: 6.0%

8.2.5 Discussion and Extensions of Simulated Model

This study does not support the hypothesis that Aurignacian populations in Early Upper Palaeolithic Europe were ethnically structured in a manner related to ornamental material culture. The spatially explicit simulation models and ABC analysis presented here, conditioned on the data presented by Vanhaeren and d'Errico (2006), indicate that there is little difference between the simple scenario of cultural identity-by-descent with modification and isolation-by-distance, and the more complex one that, in addition, invokes cultural group interaction processes that would be expected if material culture was symbolically marking ethnic group identity.

Prior to discussing the results presented above, it is important to note that any scenario considered will only be *relatively* better or worse than any other scenario considered; it is not possible to rate how good a scenario is *absolutely*.

Considering the results firstly in view of the two environmental scenarios used to condition the demography of the simulation space, we see that there is no improvement in the fit of simulated to observed data when conditioning simulations on the distribution from the Bocquet-Appel et al. study (Bocquet-Appel et al. 2005) rather than the scenario in which Europe is assumed to have a flat distribution of carrying capacities (i.e. Europe is assumed to be a flat space). Indeed, for each of the two mutation models considered, simulations in which the demography is conditioned on the latter environmental scenario generate a better fit to the observed data. Since this latter scenario is clearly not realistic, this result implies that *both* environmental scenarios used to condition the demography of the simulation space in this study are unrealistic; this is further discussed as a caveat of the current modelling framework below, along with suggestions for possible improvements.

Analysing the results now in view of the two mutation models considered, we see that, regardless of the assumed environmental scenario, data simulated under the bounded discrete mutation model generate a better fit to observed data than that simulated under the bounded stepwise mutation model. Although this result requires further investigation, it could be speculated that this suggests that, in the context at least of group interactions, there is little scaling of item similarity in material culture repertoires; little or no scaling of item similarity implies that a particular item would have been treated as either the same as or different to items already in the repertoire.

The fact that the best fits of simulated to observed data are generated by data simulated under the scenario that combines the bounded discrete mutation model and the environmental scenario in which Europe is assumed to be a flat space, and that these are a far better fit than any of the other scenarios

considered, implies that both the assumed mutation model and the assumed environmental model strongly drive the fit of simulated to observed data. Although we are cautious about interpreting the following, it is interesting to note that the assumed mutation model makes a bigger difference than the assumed environmental scenario to the fit of simulated to observed data, suggesting that continuously scaled cultural similarities were not important in distinguishing inter-group identity.

This study is a work in progress and there are several caveats, discussed below, which should be taken into consideration when interpreting our results, but these methods offer the opportunity to formally investigate whether observed material culture distributions are better explained under the assumption that ethnic structuring exists and that identities reflected in ornamental material culture influence how people interact.

It should be noted that, for the ABC approach adopted here, the number of simulations per model is relatively small and may not be enough to adequately explore the parameter space considered; for this reason, the number of simulations performed under each combination of environmental scenario and cultural variation model should be systematically increased.

The culture transmission process used in this framework assumes neutrality in that bead types are not assumed to differentially affect group fitness. A number of authors have been unable to reject neutrality using cultural transmission models (Neiman 1995; Steele et al. 2010); however, this may be due to the lack of statistical methods available to test for deviations from neutrality. Tests for deviations from neutrality have only been carried out on post-Palaeolithic datasets and have not been applied in a Palaeolithic context. However, there is certainly no a priori reason why use of different bead types should differentially affect group fitness.

As detailed in the description of the framework above, each group in the simulation deposits its material culture at the locations of the archaeological sites in the observed data (Vanhaeren and d'Errico 2006) and overwrites any existing deposits in the site if it comes within closer proximity to that site than the last group that deposited its culture there. The simulated material culture data is therefore a collection of items selected from different groups' material cultures (each of which is the result of mutation, drift and cultural interaction processes) and deposited at different points throughout the time period of interest; the process of a group depositing its culture is only dependent on the geographic distance between the group and the location of the site and deposits are made with equal probability throughout the simulation. Each site is considered to be single occupancy—only the material culture of the last group that deposited at a particular site is considered. Assuming that each site is single occupancy may be misleading since the observed data (Vanhaeren and d'Errico 2006) cannot be chronologically resolved and some sites may feature multiple layers that were deposited thousands of years apart within the period of interest. To address the inconsistency of this assumption with the cumulative aspect of the archaeological record, the depositing process could be modified such that, instead of overwriting previous deposits at a particular site, a group depositing its culture at that site would simply add its entire culture, including information on the frequency of each trait, to the existing deposits. At the end of the simulation, a number of unique items, equal to that recovered from the site as reported in the original study (Vanhaeren and d'Errico 2006), could then be selected probabilistically (using a multinomial function so that trait frequencies are taken into account) for each site, such that the probability of an item being selected is proportional to the frequency with which it is occurs in that site.

The two environmental scenarios used to condition the demography of the simulation space in this study are not realistic. In the first scenario Europe is assumed have a flat distribution of carrying capacities; this is clearly a simplistic and unrealistic view since topographic and climatic variation within the geographic region considered during the time period of interest would have had an impact on differences in habitability, and therefore the carrying capacity values, of different geographic locations at different points in time throughout the time period of interest. In the second scenario, information on estimated population densities is taken from the Bocquet-Appel et al. (2005) study to inform on carrying capacities. The reported geographic distribution and relative estimates of Upper Palaeolithic population size are an indicative starting point; however, the study itself could be considered somewhat controversial since the millennial scale climatic variation observed during the time periods that are considered is not taken into account. The geographic region during the time period of interest in the current study is characterised by a number of rapid climatic changes (Banks et al. 2008) and it is therefore unrealistic to consider the environment, and the resulting potential population densities, static for the entire duration of a simulation.

Since these environmental scenarios are unrealistic, future work could consider how the results are affected when simulations are dependent on modelled environments that take into account the climatic variability across the geographic region considered during the time period of interest. This could be achieved by using simulated Palaeoclimate data to inform on the relative carrying capacity values, and therefore potential population densities, of locations in the region of interest. Since Palaeoclimate data are available at different time points throughout the time period of interest, this approach would allow us to take into account the observed climatic variability by updating the carrying capacities in the modelled domain throughout the simulation. On way

of doing this would be to use Palaeoclimate data (Banks et al. 2008; Singarayer and Valdes 2010) to approximate Net Primary Productivity values for each location in the region of interest, following the precedent set by Eriksson et al. (2012). Net Primary Productivity provides a proxy for food availability and has been shown to be a predictor of demographic patterns in ecological studies (Binford 2001; Luck 2007) it is therefore an informative proxy for carrying capacity values, and thus potential population densities.

Group migrations could be conditioned on topographic roughness by using Topographic Roughness Index values calculated at the required resolution for the geographic area of interest using high-resolution (3 arc-sec or 90 m) elevation data (Jarvis et al. 2008). In this case, the value of the Topographic Roughness Index at a particular location would effectively scale the distance that a group can travel at that location; at locations with low values of the index (low topographic roughness) migrations would be relatively easier, while at locations with high values of the index (high topographic roughness) migrations would be relatively more difficult.

Additionally, migratory processes could be modelled as parameterised Lévy random walks, instead of as parameterised Gaussian random walks as presented above. Lévy walks are a type of random walk in which movement distances follow power-law distributions, and studies (Brown et al. 2007; Raichlen et al. 2014) looking at foraging patterns in human hunter-gatherer populations have suggested that Lévy walks are the optimal movement pattern when foraging for heterogeneously located resources (with little or no prior knowledge of resource distribution patterns). With this in mind, migratory processes in this study could be modelled as parameterised Lévy random walks, with the distance that each group traverses in a migration process selected from parameterised power-law distributions.

More generally, we have to face up to the degree of archaeological resolution we have available. Just as we cannot assume constant climatic conditions during the course of the time period considered, with climatic fluctuations that occurred during the approximately 13,000 years of the Aurignacian inevitably affecting regional population densities, we cannot necessarily assume that the aggregate data set we are dealing with represents interaction processes acting uniformly over that period; it might represent a spurious averaging of a variety of different processes. However, this is not an argument against modelling approaches; such approaches are the *only* way we can get an insight into the accumulated outcomes of iterated processes going on for hundreds or thousands of years. It is instead an argument for improving the archaeology of the time period, as well as for further comparison. The results presented here would gain further significance if they could be compared with those from the subsequent Gravettian and later cultural periods of the same region. Similarly, we may gain further insight into group interaction dynamics during the Aurignacian by comparing the results of the bead analysis (Vanhaeren and d'Errico 2006) with patterns derived from similarities and differences between lithic assemblages at the same sites.

8.3 General Discussion and Conclusions

Many fields, including archaeology, are becoming increasingly systematic and interdisciplinary through integration of traditional methods with techniques developed in other fields. Simulation modelling involves the use of theory developed for problems in physical and biological sciences and allows archaeologists to propose and test explicit hypotheses in order to address longstanding archaeological and anthropological questions. Our paper has demonstrated a novel and rigorous approach to a topic of major interest, namely the role of social structuring in archaic humans. As pointed out above, the appearance of personal ornaments has long been considered a distinctive feature differentiating Neanderthals and anatomically modern humans. While that may be the case, our results show that we have to be careful of making the further inferential step of assuming that this reflects ethno-linguistic structuring specific to anatomically modern humans. As Kuhn (2013, p. 208) points out, apparently complex large-scale phenomena can arise *"as a function of simple transmission rules operating on bounded social networks"*, thus other, and simpler, processes accounting for the observed patterning need to be considered and rejected. Simulation modelling within the Bayesian ABC framework provides a means of doing this.

As far as we are aware, the approach reported here has not been attempted when considering archaeological evidence for ethnic structuring. We fully accept that there are strengths and weaknesses to this approach, just as there are with other approaches, and these should be considered when interpreting and comparing these results to those of others. Given that there is, to the best of our knowledge, little or no representation in the literature of explicit simulation modelling approaches to questions of ethnic structuring, while interpretative approaches are well represented, we believe that this study begins to fill an important gap in the literature.

A simulation modelling approach is considerably more complex and laborious to implement compared to the interpretation of descriptive statistics or patterns in data alone. It is, however, a formal scientific approach that proposes a model with an explicit prediction of the distribution of material culture data, and tests this formally by comparing the simulated data to the observed data for validation. Taking this approach necessitates reduced models and these, by definition, will never fully describe the complexity of the true processes that shaped the material

culture data. However, the model building and testing process is not a closed one; the previous section has already indicated various ways in which aspects of the current model could potentially be improved. None-the-less, the approach adopted here is explicit and transparent and therefore less likely to be influenced by the subjective biases that guide interpretation (Gerbault et al. 2014).

Acknowledgments The authors thank Pascale Gerbault, Kevin Bryson, Bill Croft, Mark Maslin and Cynthia Beall for helpful comments and discussions. The authors also wish to thank Ken Aoki and an anonymous reviewer for constructive comments on an earlier version of this manuscript. The authors acknowledge the use of the UCL *Legion* High Performance Computing Facility, and associated support services, in the completion of this work. Mirna Kovacevic is funded by EPSRC through UCL CoMPLEX. Francesco d'Errico acknowledges the European Research Council (FP7/2007/2013, TRACSYMBOLS 249587).

Appendices

Appendix 1: Bayesian Inference and Approximate Bayesian Computation (ABC)

Bayesian inference is a branch of statistics that uses observations of particular datasets to infer the probability that a proposed hypothesis, or a parameter of that hypothesis, is true. To do this, various models with set numbers of parameters are proposed, and the posterior probability distributions of these parameters are inferred using information from prior probability distributions of the parameters and information provided by the observed data, through implementing Bayes theorem. Bayes theorem states that, given parameter (or set of parameters) θ and observed dataset D, the posterior distribution of θ, denoted $P(\theta|D)$, is proportional to the product of the probability of observing dataset D given model with parameter θ, denoted $P(D|\theta)$, and the likelihood of θ, denoted $\pi(\theta)$, which is the distribution of θ prior to any observations being made. Mathematically, this can be written as:

$$P\left(\theta\middle|D\right) \propto P\left(D\middle|\theta\right) \cdot \pi\left(\theta\right). \tag{8.14}$$

Since the explicit form of the likelihood $P(D|\theta)$ is difficult to compute in many complex problems, a family of Bayesian methods, referred to as Approximate Bayesian Computation (ABC), which do not require the likelihood function to be theoretically specified, are used (Tavare et al. 1997; Fu and Li 1997; Beaumont et al. 2002; Bertorelle et al. 2010).

In ABC techniques, a large number of datasets are simulated under a model assuming different, randomly chosen, parameter values from within prior ranges, and appropriate summary statistics are used to measure the extent to which the simulated datasets emulate the observed data. Parameter values under which the model generates datasets closest to the observed data are retained in the posterior probability distributions of the parameters.

To be able to compare the observed and simulated datasets, robust statistics that sufficiently describe the full properties of the data are used. These are called summary statistics and those developed for the current framework are discussed in detail in Appendix 3: Summary Statistics. By comparing summary statistics calculated for each simulated dataset to those for the observed data, we are able to accept to the posterior those simulations with summary statistics sufficiently close to the summary statistics for the observed dataset, referred to as the target summary statistics. The similarity δ between observed data, S, and simulated data, S', is calculated as the sum of normalised Euclidean distances of individual summary statistics:

$$\delta\left(S, S'_j\right) = \sqrt{\sum_{i=1}^{n} \frac{\left(s_i - s_{ij\prime}\right)^2}{\sigma\left(s_i\prime\right)^2}}, \tag{8.15}$$

where s and s' are values of each of the summary statistics for the observed and simulated datasets, respectively, subscript i denotes the ith of n statistics, subscript j denotes the jth of N simulations and $\sigma(s_i')$ is the standard deviation of the ith statistics over all N simulations. In performing the data analysis, we regard the ε quantile of the distribution of distances between the observed and simulated data, $\delta(S, S_j')$, as the best simulations—those generating data most similar to the observed data.

Appendix 2: Approximate Bayesian Computation (ABC) Algorithm

Let M denote the chosen model and the set of parameters of M be $\theta = (\theta_1, \ldots, \theta_m)$. Let $S = (s_1, \ldots, s_n)$ and $S' = (s_1', \ldots, s_n')$ denote the values of the summary statistics for the observed and simulated datasets, respectively. Values $S = (s_1, \ldots, s_n)$ are referred to as the target values for each of the summary statistics. The ABC algorithm is applied as follows:

1. Define a set of summary statistics that capture relevant information contained in the observed dataset.
2. Compute summary statistics values $S = (s_1, \ldots, s_n)$ for the observed dataset—these are the target values.
3. Sample parameters $\theta* = (\theta_1*, \ldots, \theta_m*)$ from an appropriate prior distribution.
4. Simulate data by using parameter $\theta*$ set with model M.
5. Compute summary statistics values $S' = (s_1', \ldots, s_n')$ for the simulated data.
6. Compute $\delta(S, S')$, where δ is an appropriately chosen distance measure.
7. For a chosen tolerance ε, retain parameter set $\theta*$ in the posterior distribution of θ if $\delta(S, S') < \varepsilon$.

8. Repeat steps 1–7 until the desired number of parameter values have been accepted to the posterior distribution.

In order for ABC methods to be effective, appropriate summary statistics that sufficiently describe the observed dataset need to be developed and appropriate choices for the distance measure, δ, and tolerance, ε, must be made.

Appendix 3: Summary Statistics

As explained previously, to be able to compare simulated and observed datasets using ABC methods, summary statistics that capture the information contained in the observed data must be developed. These should be robust statistics and should describe sufficiently the full properties of the observed dataset considered. For the current dataset, these are:

- shared information between bead types and sites, respectively
- mutual dependence between bead types and sites, respectively
- diversity in the number of occurrences of different bead types
- cultural diversity of sites as represented by the variation in the number of distinct bead types recovered from each sites
- spatial distribution of sites

For each of these statistics, we consider the values of the mean and variance in the data analysis.

Shared Information (SI)

Shared information, denoted *SI*, is a statistic that measures the extent of similarity between two variables. For measuring the shared information between bead types, *SI* is defined to be:

$$SI\left(t_i, t_j\right) = \frac{f_i f_j}{\overline{f}^2} \log \frac{r\left(t_i\right) + r\left(t_j\right)}{r\left(t_i, t_j\right)}, \tag{8.16}$$

where $r(t_i)$ and $r(t_j)$ denote the ratio of the number of occurrences of bead types i and j to the total number of sites, $r(t_i, t_j)$ is the ratio of the number of concurrent occurrence of bead types i and j to the total number of sites, f_i and f_j represent the number of sites in which bead types i and j occur, respectively, and $\overline{f}$ is the average number of times any bead type occurs over all sites. In this case, *SI* measures the similarity between pairwise bead types in terms of which sites the are present in. When two bead types never occur in the same site,

$$r\left(t_i\right) + r\left(t_j\right) = r\left(t_i, t_j\right), \text{ and} \tag{8.17}$$

$$SI\left(t_i, t_j\right) = 0. \tag{8.18}$$

A similar equation can be used to measure the shared information between sites:

$$SI\left(s_i, s_j\right) = \frac{g_i g_j}{\overline{g}^2} \log \frac{r\left(s_i\right) + r\left(s_j\right)}{r\left(s_i, s_j\right)}, \tag{8.19}$$

where $r(s_i)$ and $r(s_j)$ denote the ratio of the number of sites in which bead types i and j occur to the total number of bead types, $r(s_i, s_j)$ is the ratio of the number of sites that share bead types i and j to the total number of bead types, g_i and g_j represent the total number of bead types present in sites i and j, respectively, and $\overline{g}$ is the average number of bead types occurring per site. In this case, *SI* measures the extent of similarity between pairwise sites in terms of bead types present in those sites. Similarly to above, if two sites have no bead types in common,

$$r\left(s_i\right) + r\left(s_j\right) = r\left(s_i, s_j\right), \text{ and} \tag{8.20}$$

$$SI\left(s_i, s_j\right) = 0. \tag{8.21}$$

Mutual Information (MI)

The mutual information, *MI*, between two random variables X and Y is a measure of the mutual dependence between them. It is defined as:

$$MI\left(X;Y\right) = \sum_{y \in Y} \sum_{x \in X} p\left(x, y\right) \log \frac{p\left(x, y\right)}{p_1(x) + p_2(y)}, \tag{8.22}$$

where $p(x,y)$ denotes the joint probability of x and y (the probability of x and y occurring together), and $p_1(x)$ and $p_2(y)$ denote the marginal probabilities of x and y respectively (the probabilities of the specified values of x and y occurring).

For the observed dataset in this study, setting $X = t_i$ and $Y = t_j$, where t_i and t_j correspond to the number of occurrences of bead type i and j in all sites respectively, allows the mutual information between all pairs of bead types to be computed. Analogously, setting $X = s_i$ and $Y = s_j$, where s_i and s_j correspond to the total number of bead types present in sites i and j respectively, allows the mutual information between all pairs of sites to be computed.

In contrast to the *SI* statistic, which only examines the common presences between sites or bead types, the *MI* statistic examines both the common presences and common absences. It therefore represents the *dependence* between the pairwise vectors in question.

Mean Absolute Deviation (MAD)

The observed dataset shows large fluctuations both in the number of bead types recovered at individual sites, and

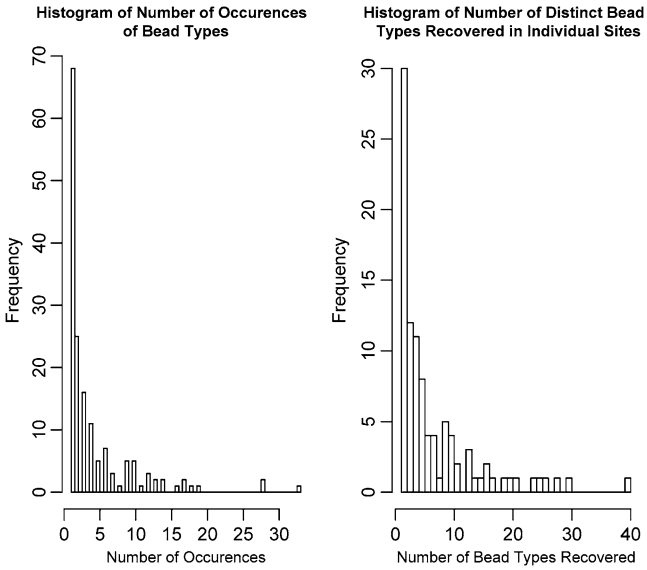


Fig. 8.3 Histograms of the number of occurrences of bead types (*left*) and number of distinct bead types recovered from individual sites (*right*) for the observed data

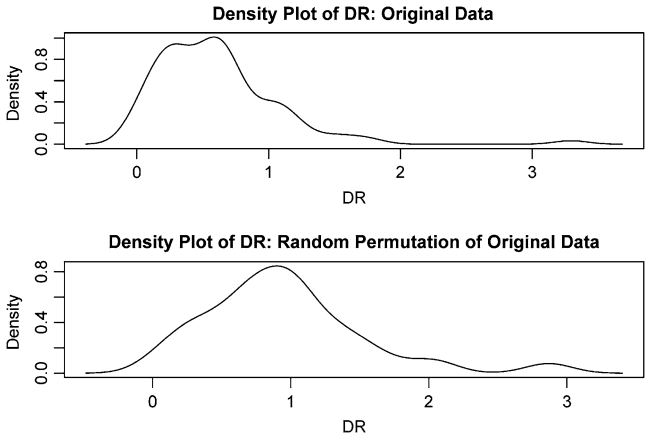


Fig. 8.4 Density plots of the *DR* statistic for the original observed data (*top*) and a random permutation of the same (*bottom*)

the number of times each particular bead type occurs, as shown in Fig. 8.3. Assuming that this is not the result of archaeological bias, these differences could be attributed to cultural wealth at sites, and the preference for particular bead types, respectively. To quantify this, the median absolute deviation statistic, *MAD*, is used. It is a measure of the variability of a random sample, and is defined to be:

$$MAD = median\left(\left|X_i - median_j\left(X_j\right)\right|\right). \tag{8.23}$$

Letting $X_i = T = \frac{f_i}{\overline{f}}$, where f_i represents the number of sites in which bead type i occurs and $\overline{f}$ is the average number of times any bead type occurs over all sites, the *MAD* statistic is a measure the variability in the number of occurrences of bead types. This can be thought of as a measure of variability in the popularity of, or preference for, bead types.

Letting $X_i = S = \frac{g_i}{\overline{g}}$, where g_i represents the total number of bead types present in site i and $\overline{g}$ is the average number of bead types occurring per site, the *MAD* statistic measures the variability in the number of beady types recovered. This can be thought of as a measure of variability in the cultural wealth recovered from sites.

Spatial Distribution of Sites (DR)

The extent to which sites share bead types may be a function of the distance between those sites. It is logical to expect that sites which are located near to each other share bead types more frequently than those which are far apart. The spatial distribution of sites can be explored by considering the average distance between sites sharing bead type i, $\overline{d}_i$, in relation to the average distance between all sites, $\overline{d}$, as follows:

$$DR_i = \frac{\overline{d}_i}{\overline{d}}. \tag{8.24}$$

DR therefore quantifies the spatial distribution of sites in terms of the shared bead types between them. Figure 8.4 shows density plots for the original observed dataset (top) and a random permutation of the same (bottom). The obvious shift to the right in the density plot of the permuted dataset implies that the distance between sites sharing a particular bead type is on average larger if bead types are randomly assigned to sites. For the original observed dataset this implies that sites which are located closer to one another on average share bead types more frequently with each other than with sites that are further away, as expected.

Appendix 4: Bayes Factors for Model Comparison

Another useful feature of the ABC approach is the ability to formally compare the performance of different models using Bayes Factors (Kass and Raftery 1995). A Bayes Factor is a summary of the evidence provided by the data in favour of one model over another. Given models M_0 and M_1, not necessarily with the same number of parameters, Bayes Factor B is given by:

$$B = \frac{P\left(M_1\middle|D\right)}{P\left(M_0\middle|D\right)} = \frac{P\left(D\middle|M_1\right)\pi\left(M_1\right)}{P\left(D\middle|M_0\right)\pi\left(M_0\right)}, \tag{8.25}$$

where $\pi(M_i)$ is the prior probability of model M_i, $P(D|\ M_i)$ is the probability of data D given model M_i and $P(M_i|D)$ is the posterior probability of the model, defined as:

$$P\ (M_i|D) = \frac{P\left(D\middle|M_i\right)\pi\ (M_i)}{P(D)}, \qquad (8.26)$$

where $P(D)$ is the unconditional marginal likelihood of the data.

This form of model comparison is independent of the parameters for each model, and instead calculates the probability of the model considering all possible parameter values. This method automatically and correctly penalises model complexity; for models with a large number of parameters there is a larger parameter space to explore and so it is more difficult to find those parameter sets that generate data similar to the observed data. Therefore, models with more parameters are penalised for the increased complexity compared to simpler models, resulting in a comparison weighted by model complexity.

References

Axelrod R (1997) The dissemination of culture—a model with local convergence and global polarization. J Confl Resolut 41(2):203–226

Banks WE, d'Errico F, Peterson AT, Kageyama M, Sima A, Sanchez-Goni MF (2008) Neanderthal extinction by competitive exclusion. PLoS One 3(12):e3972. doi:10.1371/journal.pone.0003972

Banks WE, d'Errico F, Zilhao J (2013) Human-climate interaction during the Early Upper Paleolithic: testing the hypothesis of an adaptive shift between the Proto-Aurignacian and the Early Aurignacian. J Hum Evol 64(1):39–55

Bar-Yosef O (2002) The Upper Paleolithic revolution. Annu Rev Anthropol 31:363–393. doi:10.1146/Annurev.Anthro.31.040402.085416

Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian computation in population genetics. Genetics 162(4):2025–2035

Bell AV, Richerson PJ, McElreath R (2009) Culture rather than genes provides greater scope for the evolution of large-scale human prosociality. Proc Natl Acad Sci U S A 106(42):17671–17674. doi:10.1073/pnas.0903232106

Benazzi S, Douka K, Fornai C, Bauer CC, Kullmer O, Svoboda J, Pap I, Mallegni F, Bayle P, Coquerelle M, Condemi S, Ronchitelli A, Harvati K, Weber GW (2011) Early dispersal of modern humans in Europe and implications for Neanderthal behaviour. Nature 479(7374):525–528. doi:10.1038/nature10617

Bertorelle G, Benazzo A, Mona S (2010) ABC as a flexible framework to estimate demography over space and time: some cons, many pros. Mol Ecol 19(13):2609–2625. doi:10.1111/J.1365-294x.2010.04690.X

Binford LR (2001) Constructing frames of reference: an analytical method for archaeological theory building using hunter-gatherer and environmental data sets. University of California Press, Berkeley

Bocquet-Appel JP, Demars PY, Noiret L, Dobrowsky D (2005) Estimates of Upper Palaeolithic meta-population size in Europe from archaeological data. J Archaeol Sci 32(11):1656–1668. doi:10.1016/J.Jas.2005.05.006

Brown CT, Liebovitch LS, Glendon R (2007) Levy flights in dobe ju/'hoansi foraging patterns. Hum Ecol 35(1):129–138

Buck CE (2001) Applications of the Bayesian statistical paradigm. In: Brothwell DR, Pollard AM (eds) Handbook of archaeological sciences. Wiley, Chichester, pp 695–702

Cavalli-Sforza LL, Menozzi P, Piazza A (1996) The history and geography of human genes. Abridged paperback edn. Princeton University Press, Princeton/Chichester

Clark JT, Hagemeister EM (2007) Digital discovery: exploring new frontiers in human heritage; CAA 2006: computer applications and quantitative methods in archaeology. In: Proceedings of the 34th conference, Fargo, Apr 2006. Archaeolingua, Hungary

Costopoulos A, Lake M (2010) Simulating change: archaeology into the twenty-first century, Foundations of archaeological inquiry. University of Utah Press, Salt Lake City

Eriksson A, Betti L, Friend AD, Lycett SJ, Singarayer JS, von Cramon-Taubadel N, Valdes PJ, Balloux F, Manica A (2012) Late Pleistocene climate change and the global expansion of anatomically modern humans. Proc Natl Acad Sci U S A 109(40):16089–16094

Fu YX, Li WH (1997) Estimating the age of the common ancestor of a sample of DNA sequences. Mol Biol Evol 14(2):195–199

Gerbault P, Allaby RG, Boivin N, Rudzinski A, Grimaldi IM, Pires JC, Vigueira CC, Dobney K, Gremillion KJ, Barton L, Arroyo-Kalin M, Purugganan MD, de Casas RR, Bollongino R, Burger J, Fuller DQ, Bradley DG, Balding DJ, Richerson PJ, Gilbert MTP, Larson G, Thomas MG (2014) Storytelling and story testing in domestication. Proc Natl Acad Sci U S A 111(17):6159–6164

Goldstein DB, Linares AR, Cavallisforza LL, Feldman MW (1995a) An evaluation of genetic distances for use with microsatellite loci. Genetics 139(1):463–471

Goldstein DB, Ruiz Linares A, Cavalli-Sforza LL, Feldman MW (1995b) Genetic absolute dating based on microsatellites and the origin of modern humans. Proc Natl Acad Sci U S A 92(15):6723–6727

Henrich J (2004) Demography and cultural evolution: how adaptive cultural processes can produce maladaptive losses—the Tasmanian case. Am Antiq 69(2):197–214

Higham T, Compton T, Stringer C, Jacobi R, Shapiro B, Trinkaus E, Chandler B, Groning F, Collins C, Hillson S, O'Higgins P, FitzGerald C, Fagan M (2011) The earliest evidence for anatomically modern humans in northwestern Europe. Nature 479(7374):521–524. doi:10.1038/nature10484

Higham T, Basell L, Jacobi R, Wood R, Ramsey CB, Conard NJ (2012) Testing models for the beginnings of the Aurignacian and the advent of figurative art and music: the radiocarbon chronology of Geissenklosterle. J Hum Evol 62(6):664–676. doi:10.1016/j.jhevol.2012.03.003

Hughes JK, Haywood A, Mithen SJ, Sellwood BW, Valdes PJ (2007) Investigating early hominin dispersal patterns: developing a framework for climate data integration. J Hum Evol 53(5):465–474. doi:10.1016/J.Jhevol.2006.12.011

Jarvis A, Reuter HI, Nelson A, Guevara E (2008) Hole-filled SRTM for the globe version 4. International Centre for Tropical Agriculture (CIAT); data available from the CGIAR-CSI SRTM 90 m Daab ase. (http://srtm.csi.cgiar.org)

Jones S (1997) The archaeology of ethnicity: constructing identities in the past and present. Routledge, London/New York

Kass RE, Raftery AE (1995) Bayes factors. J Am Stat Assoc 90(430):773–795

Kimura M, Ohta T (1978) Stepwise mutation model and distribution of allelic frequencies in a finite population. Proc Natl Acad Sci U S A 75(6):2868–2872

Kuhn SL (2013) Questions of complexity and scale in explanations for cultural transitions in the Pleistocene: a case study from the

Early Upper Paleolithic. J Archaeol Method Theory 20(2):194–211. doi:10.1007/s10816-012-9146-3
Kuhn SL, Stiner MC, Reese DS, Gulec E (2001) Ornaments of the earliest Upper Paleolithic: new insights from the Levant. Proc Natl Acad Sci U S A 98(13):7641–7646. doi:10.1073/pnas.121590798
Luck GW (2007) The relationships between net primary productivity, human population density and species conservation. J Biogeogr 34(2):201–212. doi:10.1111/J.1365-2699.2006.01575.X
Mellars P (2005) The impossible coincidence. A single-species model for the origins of modern human behavior in Europe. Evol Anthropol 14(1):12–27. doi:10.1002/Evan.20037
Mithen S, Reed M (2002) Stepping out: a computer simulation of hominid dispersal from Africa. J Hum Evol 43(4):433–462. doi:10.1006/Jhev.2002.0584
Neiman FD (1995) Stylistic variation in evolutionary perspective—inferences from decorative diversity and interassemblage distance in Illinois Woodland ceramic assemblages. Am Antiq 60(1):7–36
Nikitas P, Nikita E (2005) A study of hominin dispersal out of Africa using computer simulations. J Hum Evol 49(5):602–617. doi:10.1016/J.Jhevol.2005.07.001
Pinhasi R, Higham TFG, Golovanova LV, Doronichev VB (2011) Revised age of late Neanderthal occupation and the end of the Middle Paleolithic in the northern Caucasus. Proc Natl Acad Sci USA 108(21):8611–8616
Powell A, Shennan S, Thomas MG (2009) Late Pleistocene demography and the appearance of modern human behavior. Science 324(5932):1298–1301. doi:10.1126/science.1170165
Premo LS (2007) Exploratory agent-based models: towards an experimental ethnoarchaeology. In: Clark JT, Hagemeister EM (eds) Digital discovery: exploring new frontiers in human heritage; CAA 2006: computer applications and quantitative methods in archaeology: proceedings of the 34th conference, Fargo, Apr 2006. Archaeolingua, Hungary, pp 22–29
Prufer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, de Filippo C, Li H, Mallick S, Dannemann M, Fu Q, Kircher M, Kuhlwilm M, Lachmann M, Meyer M, Ongyerth M, Siebauer M, Theunert C, Tandon A, Moorjani P, Pickrell J, Mullikin JC, Vohr SH, Green RE, Hellmann I, Johnson PL, Blanche H, Cann H, Kitzman JO, Shendure J, Eichler EE, Lein ES, Bakken TE, Golovanova LV, Doronichev VB, Shunkov MV, Derevianko AP, Viola B, Slatkin M, Reich D, Kelso J, Paabo S (2014) The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505(7481):43–49. doi:10.1038/nature12886
Raichlen DA, Wood BM, Gordon AD, Mabulla AZP, Marlowe FW, Pontzer H (2014) Evidence of Levy walk foraging patterns in human hunter-gatherers. Proc Natl Acad Sci U S A 111(2):728–733
Shennan S (1989) Archaeological approaches to cultural identity, vol 10, One world archaeology. Unwin Hyman, London/Boston
Shennan S (2001) Demography and cultural innovation: a model and its implications for the emergence of modern human culture. Camb Archaeol J 11(1):5–16
Shennan S (2002) Genes, memes, and human history: Darwinian archaeology and cultural evolution. Thames & Hudson, London
Shennan S, Downey SS, Timpson A, Edinborough K, Colledge S, Kerig T, Manning K, Thomas MG (2013) Regional population collapse followed initial agriculture booms in mid-Holocene Europe. Nat Commun 4:2486. doi:10.1038/ncomms3486
Singarayer JS, Valdes PJ (2010) High-latitude climate sensitivity to ice-sheet forcing over the last 120 kyr. Quat Sci Rev 29(1–2):43–55. doi:10.1016/J.Quascirev.2009.10.011
Sinnott RW (1984) Virtues of the Haversine. Sky Telescope 68(2):159
Slatkin M (1993) Isolation by distance in equilibrium and nonequilibrium populations. Evolution 47(1):264–279
Steele J, Glatz C, Kandler A (2010) Ceramic diversity, random copying, and tests for selectivity in ceramic production. J Archaeol Sci 37(6):1348–1358. doi:10.1016/J.Jas.2009.12.039
Tavare S, Balding DJ, Griffiths RC, Donnelly P (1997) Inferring coalescence times from DNA sequence data. Genetics 145(2):505–518
Thomas MG, Stumpf MP, Harke H (2006) Evidence for an apartheid-like social structure in early Anglo-Saxon England. Proc Biol Sci/R Soc 273(1601):2651–2657. doi:10.1098/rspb.2006.3627
Tishkoff SA, Verrelli BC (2003) Patterns of human genetic diversity: implications for human evolutionary history and disease. Annu Rev Genomics Hum Genet 4:293–340. doi:10.1146/annurev.genom.4.070802.110226
Tremblay M, Vezina H (2000) New estimates of intergenerational time intervals for the calculation of age and origins of mutations. Am J Hum Genet 66(2):651–658. doi:10.1086/302770
Trinkaus E, Zilhão J (2012) Paleoanthropological implications of the Peştera cu Oase and its contents. In: Trinkaus E, Constantin S, Zilhão J (eds) Life and death at the Peştera cu Oase: a setting for modern human emergence in Europe. Oxford University Press, Oxford, pp 881–911
Valdes AM, Slatkin M, Freimer NB (1993) Allele frequencies at microsatellite loci: the stepwise mutation model revisited. Genetics 133(3):737–749
Vanhaeren M, d'Errico F (2006) Aurignacian ethno-linguistic geography of Europe revealed by personal ornaments. J Archaeol Sci 33(8):1105–1128. doi:10.1016/J.Jas.2005.11.017
Wand MP, Jones MC (1995) Kernel smoothing. Monographs on statistics and applied probability, vol 60, 1st edn. Chapman & Hall, London/New York
Wright S (1943) Isolation by distance. Genetics 28(2):114–138
Wright S (1978) Evolution and the genetics of populations: a Treatise in four volumes, volume 4 variability within and among natural populations. The University of Chicago Press, Chicago
Zilhão J (2007) The emergence of ornaments and art: an archaeological perspective on the origins of "behavioral modernity". J Archaeol Res 15(1):1–54. doi:10.1007/S10814-006-9008-1
Zilhão J, Pettitt P (2006) On the new dates for Gorham's Cave and the late survival of Iberian Neanderthals. Before Farm 3:1–9
Zilhão J, Trinkaus E, Constantin S, Milota S, Gherase M, Sarcina L, Danciu A, Rougier H, Quilès J, Rodrigo R (2007) The Peştera cu Oase people, Europe's earliest modern humans. In: Mellars PB, Boyle K, Bar-Yosef O, Stringer C (eds) Rethinking the human revolution. McDonald Institute for Archaeological Research, Cambridge, pp 249–262

Appendix C: Research Article II

Summary

This section contains the article published as:
Bains R.K., Kovacevic M., Plaster C.A., Tarekegn A., Bekele E., Bradman N.N., Thomas M.G. (2013) "Molecular diversity and population structure at the Cytochrome P40 3A5 gene in Africa" BMC Genetics 14:34

Bains et al. BMC Genetics 2013, 14:34
http://www.biomedcentral.com/1471-2156/14/34

BMC Genetics

RESEARCH ARTICLE Open Access

Molecular diversity and population structure at the Cytochrome P450 3A5 gene in Africa

Ripudaman K Bains[1*], Mirna Kovacevic[1,2], Christopher A Plaster[1], Ayele Tarekegn[3], Endashaw Bekele[3], Neil N Bradman[4] and Mark G Thomas[1,5]

Abstract

Background: Cytochrome P450 3A5 (CYP3A5) is an enzyme involved in the metabolism of many therapeutic drugs. CYP3A5 expression levels vary between individuals and populations, and this contributes to adverse clinical outcomes. Variable expression is largely attributed to four alleles, *CYP3A5*1* (expresser allele); *CYP3A5*3* (rs776746), *CYP3A5*6* (rs10264272) and *CYP3A5*7* (rs41303343) (low/non-expresser alleles). Little is known about CYP3A5 variability in Africa, a region with considerable genetic diversity. Here we used a multi-disciplinary approach to characterize *CYP3A5* variation in geographically and ethnically diverse populations from in and around Africa, and infer the evolutionary processes that have shaped patterns of diversity in this gene. We genotyped 2538 individuals from 36 diverse populations in and around Africa for common low/non-expresser *CYP3A5* alleles, and re-sequenced the *CYP3A5* gene in five Ethiopian ethnic groups. We estimated the ages of low/non-expresser *CYP3A5* alleles using a linked microsatellite and assuming a step-wise mutation model of evolution. Finally, we examined a hypothesis that CYP3A5 is important in salt retention adaptation by performing correlations with ecological data relating to aridity for the present day, 10,000 and 50,000 years ago.

Results: We estimate that ~43% of individuals within our African dataset express CYP3A5, which is lower than previous independent estimates for the region. We found significant intra-African variability in CYP3A5 expression phenotypes. Within Africa the highest frequencies of high-activity alleles were observed in equatorial and Niger-Congo speaking populations. Ethiopian allele frequencies were intermediate between those of other sub-Saharan African and non-African groups. Re-sequencing of *CYP3A5* identified few additional variants likely to affect CYP3A5 expression. We estimate the ages of *CYP3A5*3* as ~76,400 years and *CYP3A5*6* as ~218,400 years. Finally we report that global CYP3A5 expression levels correlated significantly with aridity measures for 10,000 [Spearmann's Rho= −0.465, p=0.004] and 50,000 years ago [Spearmann's Rho= −0.379, p=0.02].

Conclusions: Significant intra-African diversity at the *CYP3A5* gene is likely to contribute to multiple pharmacogenetic profiles across the continent. Significant correlations between CYP3A5 expression phenotypes and aridity data are consistent with a hypothesis that the enzyme is important in salt-retention adaptation.

Keywords: Cytochrome P450 3A5, Africa, Population genetics, Gene-environment correlations, Pharmacogenetics

Background

One of the main goals of the genomics revolution has been to characterize diversity within indigenous populations, which have traditionally been under-represented in research. The availability of genomic data is enabling researchers to identify how and why genomic variation affects individual and population differences in clinical outcomes following pharmaceutical drug administration. Additionally, evolutionary and demographic processes which have shaped population variation at clinically relevant regions of the human genome are now being determined. Studies of genes encoding drug metabolizing enzymes, such as the Cytochrome P450 (CYP450) superfamily have identified variation which affects the safety and efficacy of therapeutic drugs. However little is known about intra-African variation at these loci. Africa is heavily burdened with common and infectious diseases [1], which are treated with multiple drugs. Studies

* Correspondence: r.bains@ucl.ac.uk
[1]Research Department of Genetics, Evolution and Environment, University College London, Darwin Building, Gower Street, London WC1E 6BT, UK
Full list of author information is available at the end of the article

of intra-African variation at genes encoding drug metabolizing enzymes are likely to be beneficial to clinicians, geneticists and researchers within the emerging field of evolutionary medicine [2]. They are also likely to have great potential for minimizing the risk of adverse clinical outcomes in patients with recent African ancestry [3].

CYP3A enzymes, a sub-family of the CYP450 superfamily, are responsible for the phase I hepatic and intestinal metabolism of a wide spectrum of endogenous and xenobiotic compounds [4]. The two most clinically relevant CYP3A enzymes are CYP3A4 and CYP3A5, which together are involved in the metabolism of ~50% of all therapeutic drugs [5]. Because of the wide substrate range, some functional variation in *CYP3A* genes is associated with individual and population differences in pharmacogenetic profiles [6], adverse clinical outcomes [7], and elevated predisposition to diseases [8,9].

There is considerable inter-ethnic variability in CYP3A5 expression levels [10]. Individuals tend to express CYP3A5 at high concentrations (21-202 pmol/mg) or have significantly reduced, often undetectable, protein levels (<21 pmol/mg) [11-13]. Variability in protein expression is largely attributed to four *CYP3A5* alleles; *CYP3A5*1*, an expresser allele, and the low/non-expresser *CYP3A5*3, CYP3A5*6* and *CYP3A5*7* alleles [13,14]. Studies have reported that the highest frequencies of high-activity alleles are found in populations with recent African ancestry [15,16]. *CYP3A5*3* is the main determinant of CYP3A5 expression levels in populations outside Africa [10]. The *CYP3A5*6* and *CYP3A5*7* alleles are observed almost exclusively in individuals with recent African ancestry [13-16], although *CYP3A5*6* has been observed at low frequency in a sample of individuals from Los Angeles with Mexican ancestry, genotyped as part of the HapMap consortium. *CYP3A5*7* has been observed at a frequency of 3% in ethnic Koreans [17]. There is some uncertainty over the functionality of the *CYP3A5*6* mutation. Its effect on protein expression was reported in 2001 [13]. One of two cDNA products isolated from three *CYP3A5*1/CYP3A5*6* heterozygotes did not contain the sequence for exon 7. Subsequent western blot analyses of liver samples from two *CYP3A5*1/CYP3A5*6* heterozygotes found significantly lower protein levels than in *CYP3A5*1* homozygotes. It has been proposed that *CYP3A5*6* creates an aberrant splicing pathway [13], however this has not been confirmed experimentally. Although data presented by Kuehl *et al.* suggest that CYP3A5 expression levels in *CYP3A5*6* carriers are lower than in *CYP3A5*1* homozygotes, in the absence of expression analysis and more extensive *in vivo* and *in vitro* data we considered it prudent to allow for the possibility that in at least some individuals *CYP3A5*6* is expressed. Unlike the *CYP3A5*3* and *CYP3A5*7* mutations, the association between *CYP3A5*6* and clinical outcomes is not completely certain. A study examining the association between *CYP3A* genotypes and the metabolism of midazolam found a significant association between the metabolism of the drug and the presence of the *CYP3A5*3* allele, but not the *CYP3A5*6* allele [18]. However, an independent study of Japanese breast cancer patients found that tumor sizes were significantly higher in women who carried the *CYP3A5*6* allele [19]. Given this uncertainty we present analyses that assume both that the *CYP3A5*6* allele does, and does not affect protein expression and function.

A previous study reported that elevated *CYP3A5*3* frequencies are positively correlated with increased geographic distance from the equator [20]. There is a latitudinal cline in the frequencies of alleles involved in heat adaptation, and consequently hypertension susceptibility [21]. A strong positive correlation is observed between latitude and functionally important variants of genes implicated in salt-sensitive hypertension, by regulating cardiovascular reactivity and volume avidity, such as angiotensinogen (*AGT*), G protein β3 subunit (*GNB3*), and epithelial sodium channel γ (*ENaCγ*) [21]. CYP3A5 is involved in the metabolism of renal cortisol to 6-β -hydroxycortisol, a key regulator of renal sodium transport, and immune responses which cause inflammation [22]. It has been proposed that the expresser *CYP3A5*1* allele provides a selective advantage in equatorial populations due to the role of CYP3A5 in salt retention and the reabsorption of water [13,20]. Conversely, elevated *CYP3A5*1* frequencies are hypothesized to be detrimental and are associated with elevated risk of salt-sensitive hypertension in non-equatorial populations [8,23,24]. The *CYP3A5* gene region has high frequencies of derived, functional alleles [25], and substantial population differentiation in the frequencies of the *CYP3A5*3* allele when compared to neutral markers, as measured by weighted F_{ST} tests, [26]. This suggests that low/non-expression of CYP3A5 may be adaptive in non-equatorial populations.

Although CYP3A5 expression in Africa is likely to be highly variable, few previous studies have characterized intra-African diversity in *CYP3A5* and other clinically relevant genes. High levels of genetic diversity are observed within the continent compared to other geographic regions, and this is consistent with a recent African origin model of human evolution [27]. East Africa is a particularly diverse region of the continent. Reports have shown a gradual reduction in genetic diversity with increased geographic distance from Ethiopia [28-30] indicating that the region is one of the most genetically diverse in the world. Studies of functional variation in clinically relevant genes have found significant inter-ethnic differences within Ethiopia and between Ethiopian and other African populations [31-33]. These data highlight the potential that focused genetic studies of clinically relevant variation within

Ethiopian populations have for understanding intra-African genetic diversity.

Within this study we have focused on characterizing *CYP3A5* variation in multiple geographically and ethnically diverse populations sampled from in and around Africa. We focused on determining population structure at this locus, and identified considerable population structuring within Africa. These results suggest that there are likely to be multiple pharmacogenetic profiles across Africa which could affect the safety and efficacy of many therapeutic drugs which are CYP3A5 substrates. Additionally, we report correlations between CYP3A5 expression phenotypes and aridity data for 10,000 and 50,000 years ago, consistent with a previous hypothesis that the enzyme is involved in salt retention/heat adaptation. This suggests that global variability in expression phenotypes may have occurred as a result of selective pressures on the gene.

Results

The prevalence of clinically relevant *CYP3A5* alleles in Africa

We genotyped 2245 individuals from 32 geographically and ethnically diverse African populations for common clinically relevant *CYP3A5* alleles. An additional 293 individuals from four non African populations from Europe and the Arabian Peninsula were also genotyped to permit comparisons of African diversity in a global context (Table 1). Prior to our study, the distribution of clinically relevant *CYP3A5* alleles across Africa, and relative to non-African populations, was unknown. We identified *CYP3A5*1, CYP3A5*3* and *CYP3A5*6* in all genotyped African population samples (allele frequency ranges: 4-81%, 4-81% and 4-33% respectively). *CYP3A5*7* was confined almost exclusively to Niger-Congo speaking samples (range: 0-22%). The distribution of *CYP3A5* alleles is structured by major language family and geographic region, as evidenced by Analysis of Molecular Variance [$P<0.0001$ for both variables]. Pearson's χ^2 tests were carried out to examine within-region differences. Considerable heterogeneity was observed in East Africa [$\chi^2=157.69$, d.f.=21, $p<0.0001$] and North Africa [$\chi^2=37.61$, d.f.=9, $p<0.01$] but not in any other geographic region. The genotyped loci are in complete LD ($D'=1$, $p<0.0001$), except between the *CYP3A5*6* and *CYP3A5*1/*3* loci ($D'=0.96$, $p<0.0001$). A low frequency recombinant haplotype was observed in 10 heterozygotes explaining why D' between *CYP3A5*1* and *CYP3A5*6* is not equal to 1. Haplotype analysis found that the low/non-expresser *CYP3A5* alleles occur predominantly on independent haplotype backgrounds (Figure 1 and Additional file 1 Table S1) suggesting that their convergent effects on CYP3A5 expression are independent. A significant correlation between pairwise genetic (F_{ST}) and geographic distances (kilometers) was observed using a Mantel test when all populations genotyped in this study ($n=36$) were analyzed [Mantel r statistic=0.228, $p<0.0001$].

The geographic and ethnic distributions of low-, intermediate- and high-expression phenotypes, based on haplotype frequencies were inferred. Expresser phenotypes were inferred assuming that *CYP3A5*6* does and does not cause a low/non-expression phenotype (Additional file 2 Figure S1 and Additional file 3 Figure S2 respectively). The distributions in both Figures show that the highest frequencies of high-activity phenotypes are in equatorial regions of Africa, and Ethiopia has the highest within country inter-ethnic diversity, which is driven by differences between the Anuak and other Ethiopian groups.

Correlations between ecological variables and inferred CYP3A5 *expression phenotypes*

A previous study reported a strong positive correlation between *CYP3A5*3* allele frequencies and latitude [20]. Latitude is a correlate of multiple ecological variables that are associated with functional markers of genes involved in heat adaptation [21]. We tested for correlations between frequencies of low/non-expresser *CYP3A5* alleles, and inferred expresser phenotypes, with latitude and the ecological variables; temperature and precipitation (Table 2). Additionally, we tested for correlations with aridity indices calculated from temperature and precipitation data using the de Martonne aridity index [34]. This enabled us to consider the combined effect of temperature and precipitation on CYP3A5 phenotypes. Correlations were estimated using ecological data for the present day, and inferred for 10, 000 years ago (Holocene) and 50,000 years ago (Late Pleistocene) (http://badc.nerc.ac.uk/home/index.html). Correlations were performed assuming that *CYP3A5*6* is a low/non-expresser allele, and that it is a neutral allele.

Latitude correlated significantly with CYP3A5 expression in Africa [Spearmann's Rho= −0.472, $p=0.004$], the correlation remained significant when considering north [Spearmann's Rho= −0.659, $p<0.0001$] and south latitude [Spearmann's Rho= −0.701, $p<0.0001$] populations separately. Across a global cohort (87 populations) which, included published genotyping data [20] and where *CYP3A5*3* alone is considered to predict CYP3A5 expression levels, a significant correlation between latitude and frequencies of this allele was seen only for north latitude populations [Spearmann's Rho= 0.666, $p<0.0001$], but not south [Spearmann's Rho= 0.066, $p=0.759$]. No significant correlation was observed between aridity values for the present day and expresser phenotypes when *CYP3A5*6* was considered a low/non expresser allele [Spearmann's Rho= −0.185, $p=0.279$] or a neutral allele [Spearmann's Rho= −0.0288, $p=0.868$]. Expresser phenotypes correlated significantly with aridity values from the Holocene [Spearmann's Rho= −0.465, $p=0.004$] and Late Pleistocene [Spearmann's Rho= −0.379, $p=0.02$] when *CYP3A5*6*

was considered as a low/non-expresser mutation. We subsequently examined independent correlations between expresser allele frequencies and temperature and precipitation. We found significant correlations between expresser allele frequencies and temperature for every time period, both when *CYP3A5*6* was considered to be a low/non-expresser mutation and a neutral allele ($p<0.0001$ for every correlation, see Table 2). No significant correlation was observed between precipitation values and expresser allele frequencies.

We subsequently examined the correlations between present day ecological data and expresser allele frequencies, while controlling for geographic distances between populations, using partial Mantel tests. For each correlation *CYP3A5*6* was assumed to be a low/non-expresser mutation. We found that the correlation between CYP3A5 expresser alleles and temperature remained significant when controlling for geographic proximity between populations [Mantel r statistic=0.398, p=0.02]. However the correlation with latitude was no longer significant [Mantel r statistic=0.202, p=0.05].

CYP3A5 variation observed in Ethiopia

Previous studies of genetic variation in drug metabolizing enzymes have identified considerable inter-ethnic diversity within Ethiopia and between Ethiopian and other African populations [31-33]. The results from our geographic survey of clinically relevant *CYP3A5* variants also indicated that there is considerable heterogeneity within Ethiopia, and between Ethiopia and other African populations. We performed a re-sequencing survey of the *CYP3A5* gene in five Ethiopian populations to characterize *CYP3A5* diversity in greater detail.

We observed significant inter-ethnic diversity in *CYP3A5* allele frequencies in Ethiopia. To identify additional variation and elucidate intra-Ethiopian population structure we re-sequenced an 8063bp region of *CYP3A5,* which included the *CYP3A5* promoter, exons and exon-flanking introns, in five Ethiopian populations. 51 polymorphic sites were identified (Table 3). Nine (17.6%) were exonic and, 3 out of 5 (6%) identified non-synonymous polymorphisms were predicted to adversely alter protein function. No significant difference in the proportion of synonymous or non-synonymous variation was identified by a codon-based Z-test [35] (Z=0.961 and p=0.169). The proportion of amino acid changes that we observed at the *CYP3A5* gene (5 changes/502 codons= ~1%) is higher than previously reported for 103 protein-coding genes (147 changes/26,999 codons=0.56%) [36], although the differences are not significant [paired t test, t=1.01, d.f.=1, p=0.50]. We did not identify any variants in experimentally established transcription factor binding sites [37,38]. Eight of the nine identified promoter variants occurred in nucleotide positions that are highly conserved in primates (i.e. where the allele is the same in all primate species), and bioinformatic analyses predicted that four out of nine may affect transcription factor binding. Of all identified polymorphisms – predicted and previously reported to affect CYP3A5 expression and activity (n=10) – 4 (2 promoter, *CYP3A5*3* and *CYP3A5*6*) occurred at frequencies over 1%. The highest frequency variants identified were *CYP3A5*3, CYP3A5*6* and the non-functional variant rs15524, which is found in high LD with *CYP3A5*3* [39].

Ethiopian *CYP3A5* variation in the context of other geographic populations

We analyzed the Ethiopian re-sequencing data along with those previously reported for three ethnically diverse populations from the Corriell Repositories to analyze the data in a global context [20] (Table 4). The results of the Hudson-Kreitman-Aguadé (HKA) test [40], comparing intra- and inter-species *CYP3A5* diversity, was not significant (p=0.6346). Tajima's D, Li's D^* and F^*, Fu and Li's F and D (using chimpanzee sequence to establish ancestral states), and Fu's F_S all indicated a skew towards rare variants in every population, which is consistent with general human population growth or positive selection. Fu and Li's D^* and F^* reported a significant departure from neutrality for both Europeans and the Anuak, although significance was only reached for Europeans following Bonferonni correction (8 tests). Fu and Li's F_S reported a significant departure from neutrality for 7 of the 8 populations after Bonferonni correction. Strobeck's S results were consistent with Fu's F_S, as expected. The results of the H test, used to assess whether there is an excess of high frequency derived variants [41], were not significant in any population ($p>0.05$), however nucleotide diversity at *CYP3A5* is low and this may be affecting the tests.

72 haplotypes were inferred from allelic data for all 8 Ethiopian population samples, 33 (45.8%) containing *CYP3A5*1,* 29 (40.3%) containing *CYP3A5*3,* 7 (9.7%) containing *CYP3A5*6*, 1 (1.4%) containing *CYP3A5*7*, and 2 (2.8%) containing both *CYP3A5*3* and *CYP3A5*6* (Additional file 4 Figures S3a and b). LD across the gene is high. A phylogeny, based on network analysis of the haplotype data, is presented in Figure 2. 98% of European and 83% of Han Chinese haplotypes contain the *CYP3A5*3* allele, as do ~64% of Afar haplotypes and ~67% in both the Amhara and Oromo. Gene diversity is highest in African Americans (0.963 ± 0.02) and lowest in Europeans (0.589 ± 0.08). The *CYP3A5*1* haplogroup is significantly more diverse than the other haplogroups (0.921 ± 0.01) ($p<0.0001$ for every comparison). Population differentiation was measured by pairwise F_{ST} (Table 5). The Afar, Amhara and Oromo are intermediate between individuals with recent African ancestry and Han Chinese and European groups. We placed population structure seen at the *CYP3A5* gene in a wider genomic

Table 1 Genotype and allele frequencies and tests for deviation from Hardy-Weinberg Equilibrium (χ^2 *p*-values given)

Region	Country	Population	*CYP3A5*1/CYP3A5*3*						*CYP3A5*6*						*CYP3A5*7*					
			AA	AG	GG	Total	G [*3]	HWE	GG	GA	AA	Total	A [*6]	HWE	–/–	-/T	T/T	Total	T [*7]	HWE
Europe	Armenia	Southern Armenians	0	10	90	100	0.95	1.00	100	0	0	100	0.00	N/A	100	0	0	100	0.00	N/A
	Turkey	Anatolian Turks	2	10	62	74	0.91	0.11	74	0	0	74	0.00	N/A	74	0	0	74	0.00	N/A
Arabian	Yemen	Yemeni from Hadramaut	2	21	59	82	0.85	1.00	77	5	0	82	0.03	1.00	80	2	0	82	0.01	1.00
Peninsula		Yemeni from Sena and Msila	7	17	13	37	0.58	0.74	29	7	1	37	0.12	0.42	35	2	0	37	0.03	1.00
North Africa	Algeria	Northern Algerians	9	42	108	159	0.81	0.12	146	15	0	161	0.05	1.00	159	2	0	161	0.01	1.00
	Morocco	Berbers	3	28	54	85	0.80	1.00	79	7	0	86	0.04	1.00	85	1	0	86	0.01	1.00
	Sudan	Northern Sudanese	24	58	51	133	0.60	0.29	104	28	0	132	0.11	0.36	135	1	0	136	0.00	1.00
		Sudanese from Kordofan	11	11	8	30	0.45	0.16	19	10	1	30	0.20	1.00	29	1	0	30	0.02	N/A
East Africa	Ethiopia	Afar	10	31	32	73	0.65	0.61	47	26	0	73	0.18	0.11	73	0	0	73	0.00	N/A
		Amhara	14	22	40	76	0.67	0.004	55	19	2	76	0.15	0.67	76	0	0	76	0.00	N/A
		Anuak	38	32	6	76	0.29	1.00	44	25	7	76	0.26	0.23	75	1	0	76	0.01	1.00
		Maale	20	36	19	75	0.49	0.82	53	22	0	75	0.15	0.34	74	1	0	75	0.01	1.00
		Oromo	12	28	34	74	0.65	0.20	55	19	1	75	0.14	1.00	75	0	0	75	0.00	N/A
	Republic of South Sudan	Southern Sudanese	74	42	9	125	0.24	0.46	58	50	15	123	0.33	0.42	117	8	0	125	0.03	1.00
	Tanzania	Chagga	28	18	4	50	0.26	0.71	36	14	0	50	0.14	0.57	41	9	0	50	0.09	1.00
	Uganda	Bantu speakers from Ssese	36	3	0	39	0.04	1.00	22	17	0	39	0.22	0.16	23	16	0	39	0.21	0.31
West Africa	Ghana	Asante	27	8	0	35	0.11	1.00	20	13	1	34	0.22	1.00	29	5	0	34	0.07	1.00
		Bulsa	58	29	3	90	0.19	1.00	61	28	0	89	0.16	0.11	69	19	2	90	0.13	0.62
		Kasena	28	17	2	47	0.22	1.00	31	16	0	47	0.17	0.32	35	12	0	47	0.13	1.00
	Senegal	Manjak	57	29	4	90	0.21	1.00	59	24	9	92	0.23	0.02	81	13	0	94	0.07	1.00
		Wolof	55	31	8	94	0.25	0.27	58	31	1	90	0.18	0.29	78	15	1	94	0.09	0.55
West Central	Cameroon	Kotoko	18	21	0	39	0.27	0.04	23	16	1	40	0.23	0.65	36	4	0	40	0.05	1.00
Africa		Shewa Arabs	26	31	12	69	0.40	0.62	42	24	3	69	0.22	1.00	60	9	0	69	0.07	1.00
		Mayo Darle	66	38	13	117	0.27	0.06	71	33	13	117	0.25	0.01	102	15	0	117	0.06	1.00
		Somie, Cameroonian Grassfields	36	28	1	65	0.23	0.16	44	19	2	65	0.18	1.00	52	13	0	65	0.10	1.00
	Congo	Congolese from Brazzaville	35	18	2	55	0.20	1.00	43	11	1	55	0.12	0.55	45	10	0	55	0.09	1.00
	Nigeria	Igbo	64	23	0	87	0.13	0.35	60	24	4	88	0.18	0.47	73	12	2	87	0.09	0.14

Table 1 Genotype and allele frequencies and tests for deviation from Hardy-Weinberg Equilibrium (χ^2 *p*-values given) *(Continued)*

South East	Malawi	Chewa	66	25	1	92	0.15	1.00	66	23	3	92	0.16	0.69	60	31	0	91	0.17	0.06
Africa		Lomwe	13	4	1	18	0.17	N/A	10	8	0	18	0.22	N/A	14	4	0	18	0.11	N/A
		Ngoni	15	2	1	18	0.11	N/A	9	6	3	18	0.33	N/A	16	2	0	18	0.06	N/A
		Tumbuka	44	18	0	62	0.15	0.34	40	17	5	62	0.22	0.14	45	17	0	62	0.14	0.59
		Yao	37	18	1	56	0.18	0.67	43	12	1	56	0.13	1.00	46	10	0	56	0.09	1.00
	Mozambique	Sena	58	21	3	82	0.16	0.44	51	28	5	84	0.23	0.75	59	25	1	85	0.16	0.68
	South Africa	Bantu speakers	22	17	2	41	0.26	1.00	29	9	3	41	0.18	0.10	34	4	2	40	0.10	0.03
	Zimbabwe	Lemba	17	6	0	23	0.13	1.00	13	10	1	24	0.25	1.00	17	7	0	24	0.15	1.00
		Zimbabweans from Mposi	36	7	4	47	0.16	0.008	36	10	3	49	0.16	0.09	34	16	2	52	0.19	1.00

HWE could not be calculated for the Lomwe and Ngoni as both populations had fewer than 50 chromosomes meaning that the test had insufficient power. No population deviated from Hardy-Weinberg equilibrium, following Bonferonni correction (for *CYP3A5*3*: adjusted *p* value = 0.00139; correction for 36 tests, for *CYP3A5*6*: adjusted *p* value=0.0015; correction for 34 tests, for *CYP3A5*7*: adjusted *p* value=0.0017; correction for 30 tests). Deviations from HWE cannot be calculated for monomorphic loci: labeled "N/A" on the Table. "Total" refers to the number of individuals, from a given population, successfully genotyped at each locus. Population refers to the grouping of individuals either by self-declared ethnicity or geography/place collected.

context by analyzing intra-Ethiopian differentiation at markers on the non-recombining regions of the Y chromosome (NRY) and the mitochondrial genome (hypervariable region 1 [HVS1] and coding region SNPs) [42]. We compared Ethiopian NRY and HVS1 genotypes with data for 92 Fars from Iran, 95 Nigerian Igbo, 126 Greek-Cypriots and 60 Halfawi from the Republic of Sudan. The Anuak are outliers compared to the other Ethiopian populations (data not shown), consistent with genome wide-markers [43] and what we report for *CYP3A5*. Intra-Ethiopian population structure at the *CYP3A5* gene is also consistent with that seen at other drug metabolizing genes *CYP1A2* [31], and *UGT1A1* [32].

Estimating the age of clinically relevant *CYP3A5* alleles

The age of an allele is the time since it arose by mutation [44,45]. Estimating the ages of *CYP3A5* alleles may help to identify specific demographic processes which have affected inter-population differences in allele frequencies, or identify an important role for natural selection in selecting for specific alleles [44]. Under the stepwise mutation model of microsatellite evolution, and assuming no recombination, we estimated the time to the most recent ancestor (TMRCA) of the *CYP3A5*3* mutation to be 2388 generations (95% confidence intervals [C.I.]: 1797–3211) and *CYP3A5*6* to be 6825 generations (95% C.I.: 3086–11,975). Assuming that a generation is 32 years [46], the estimated age of the *CYP3A5*3* mutation is ~76,416 years (95% C.I. 57,504-102,752 years) and *CYP3A5*6* is 218,400 years (95% C.I. 98,752-383,200 years) (Table 6). Our estimates of the age of *CYP3A5*3* is consistent with its presence within and outside of Africa. The distribution of the *CYP3A5*6* allele shows some similarity to that of *FMO2*1*, an allele of the gene encoding the drug metabolizing enzyme FMO2 [33]. *FM02*1* occurs at similar frequencies across Africa and is not found at high frequencies outside of the continent. The estimated age of *FMO2*1* is 502,404 years (95% C.I. 154,790–1,041,243 years) based on a coalescent simulation [47] and using data from populations re-sequenced as part of the NIEHS SNPs database (http://egp.gs.washington.edu/). The age estimates of both the *CYP3A5*6* and *FMO2*1* alleles predate estimates of the range-expansion of modern humans out of Africa.

Discussion

We performed an extensive geographic survey of clinically relevant *CYP3A5* alleles in a large African cohort and found highly variable frequencies of the ancestral *CYP3A5*1* allele (9-96%) across the continent. We estimate that ~43% of individuals within our African dataset express CYP3A5, which is much lower than all other previous estimates for the continent (between 55-95%) [15,16]. The classification of *CYP3A5* alleles as expresser or low/non-expresser will affect estimates of expresser frequencies in Africa. *In vitro* studies of CYP3A5 expression levels in *CYP3A5*6* homozygotes are needed to establish the effect of the mutation on protein expression. The results from such studies may alter the classification of *CYP3A5*6* as a clinically relevant *CYP3A5* allele, and mean that CYP3A5 protein expression levels across Africa are likely to be consistent with those presented in Additional file 3 Figure S2 . Our estimates of the proportion of CYP3A5 expressers differ across Africa, consistent with the Sahara acting as a barrier to gene flow [48,49]. Additionally, we estimate that the proportion of CYP3A5 expressers in East Africa (~36%) is lower than in other regions of sub-Saharan Africa (~45%), and report considerable heterogeneity among Ethiopian ethnic groups (17-54%). We found that the highest frequencies of inferred high-activity phenotypes were seen in equatorial and Niger-Congo speaking populations.

From the geographic survey we observed that the Ethiopian allele frequencies are intermediate between sub-Saharan African and Eurasian groups [50]. Our study has extended previous work on *CYP3A5* in Ethiopia [51] by accounting for, and identifying, considerable inter-ethnic variability within the country. *CYP3A5* haplotype diversity and structure in the Afar, Amhara and Oromo were characteristic of that seen in European Caucasians and Han Chinese individuals. There is a known Arabian contribution to Ethiopian ancestry as a result of migration of Semitic groups into the region, which has influenced genetic diversity [48,52]. We further examined intra-Ethiopian diversity at mitochondrial and Y-chromosome genetic markers and found that the Anuak were outliers. This suggests that the intra-Ethiopian diversity we observed can be explained by Arabian admixture in the Afar, Amhara and Oromo, rather than differential selection pressures on *CYP3A5*.

Considerable intra-African population structuring at the *CYP3A5* gene suggests that there are likely to be multiple pharmacogenetics profiles for key drugs used across the continent, including many used in the treatment and control of malaria [53] and HIV-1 [54]. We identified significant differences between Ethiopians and other sub-Saharan African populations, and intra-Ethiopian diversity, at the *CYP3A5* gene. The results from our study suggest that East Africans are likely to be distinct from a wider cohort of African patients, and that there are likely to be inter-ethnic differences within East Africa. The results from large surveys [32,33], including our study, emphasize the importance of including sub-Saharan African populations in pharmacogenetics research; over 90% of the global disease burden is found in developing countries [55,56]. An appreciable number of the diseases found within the region are treated with CYP3A5 substrates [5,57] at doses optimized for patients with recent European ancestry [26]. Larger and

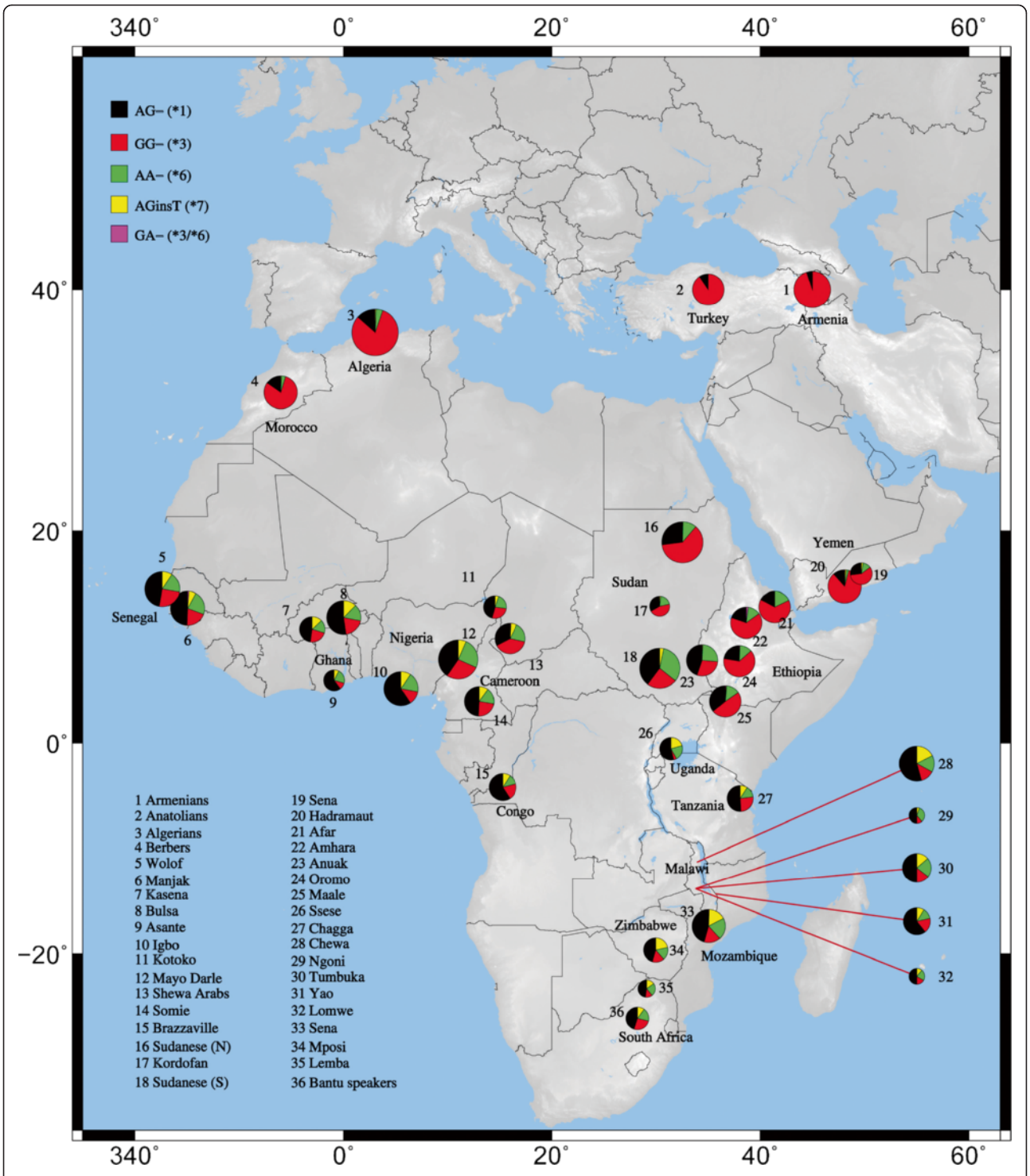


Figure 1 The distribution of the five inferred *CYP3A5* haplotypes across the dataset. The size of each circle is proportional to the number of individuals sampled from a given population (see Additional file 1 Table S1). The allele combinations at all three loci are given in the key. The alleles in brackets define the inferred haplotype. N.B. the recombinant *CYP3A5* haplotype 5: **3/*6*, is observed at low frequency in the dataset.

Table 2 Correlation analyses, between ecological variables and *CYP3A5* allelic and inferred expression data

Time period	Ecological variable	*N*=87		*N*=36											
		*CYP3A5*3*		*CYP3A5*6*		*CYP3A5*7*		High expresser allele (assuming *CYP3A5*6* is a low/non-expresser allele)		Low expresser allele (assuming *CYP3A5*6* is a low/non-expresser allele)		High expresser allele (assuming *CYP3A5*6* is not a low/non-expresser allele)		Low expresser allele (assuming *CYP3A5*6* is not a low/non-expresser allele)	
		Rho	*P*-value	Rho	*P*-value	Rho	*P*-value	Rho	*P*-value	Rho	*P*-value	Rho	*P*-value	Rho	*P*-value
	Latitude	0.706	**<0.0001**	−0.331	**0.048**	−0.177	0.303	−0.472	**0.004**	0.472	**0.004**	−0.416	**0.012**	0.416	**0.012**
	North latitude	0.666	**<0.0001**	−0.621	**0.001**	−0.410	**0.047**	−0.659	**<0.0001**	0.659	**<0.0001**	−0.620	**0.001**	0.620	**0.001**
	South latitude	0.066	0.759	0.318	0.130	0.122	0.571	−0.701	**<0.0001**	0.701	**<0.0001**	−0.370	0.075	0.370	0.075
	Temperature	−0.664	**<0.0001**	0.268	0.114	0.494	**0.002**	0.655	**<0.0001**	−0.655	**<0.0001**	0.627	**<0.0001**	−0.627	**<0.0001**
Present Day	**Precipitation**	−0.129	0.232	−0.290	0.867	−0.150	0.384	−0.028	0.869	0.028	0.869	0.113	0.511	−0.113	0.511
	Aridity	0.286	**0.007**	−0.201	0.24	−0.267	0.116	−0.185	0.279	0.185	0.279	−0.029	0.868	0.029	0.868
	Temperature	−0.597	**<0.0001**	0.216	0.207	0.342	**0.041**	0.560	**0.0004**	−0.560	**0.0003**	0.635	**<0.0001**	−0.635	**<0.0001**
Holocene	**Precipitation**	0.072	0.510	−0.235	0.167	−0.522	**0.001**	−0.381	**0.022**	0.381	**0.022**	−0.190	0.266	0.190	0.266
	Aridity	0.471	**<0.0001**	−0.344	**0.04**	−0.575	**0.0002**	−0.465	**0.004**	0.465	**0.004**	−0.293	0.083	0.293	0.0832
	Temperature	−0.644	**<0.0001**	0.297	0.079	0.608	**<0.0001**	0.649	**<0.0001**	−0.649	**<0.0001**	0.641	**<0.0001**	−0.641	**<0.0001**
Late Pleistocene	**Precipitation**	0.160	0.139	−0.238	0.163	−0.353	**0.035**	−0.204	0.233	0.204	0.233	−0.023	0.892	0.023	0.892
	Aridity	0.532	**<0.0001**	−0.436	**0.008**	−0.480	**0.003**	−0.379	**0.026**	0.379	**0.023**	−0.211	0.216	0.211	0.216

For analyses with inferred CYP3A5 expression phenotypes high-, intermediate- and low- expression diplotypes were counted as genotypes and the frequencies of expresser and low/non-expresser alleles calculated. For analyses with phenotypes, *CYP3A5*6* was considered to cause low/non-expression and to have no effect on CYP3A5 expression. Significant *p*-values, at the 5% level, are shown in bold. Rho indicates Spearmann's Rho. "*N*" refers to the number of populations analyzed for each *CYP3A5* allele. For *CYP3A5*3* frequencies, African data were combined with those previously reported [20]. For *CYP3A5*6* and *CYP3A5*7* correlations, only African data genotyped for this study were tested. North latitude and south latitude correlations were only performed with populations genotyped for this study.

Table 3 All polymorphic sites identified in an 8063bp *CYP3A5* region re-sequenced in five Ethiopian populations

					Afar		Amhara		Anuak		Maale		Oromo		Total	
Region of *CYP3A5*	**Position on chromosome 7**	**Position relative to the translation initiation codon (A of ATG is +1)**	**dbSNP database refSNP ID**	**Effect**	***f***	**n**	***f***	**n**	***f***	**n**	***f***	**n**	***f***	**n**	***f***	**n**
Promoter	99278314	−795 T>A	rs3823812		0.00	3	0.00	3	0.01	4	0.01	10	0.01	5	0.0331	25
Promoter	99278267	−748 C>G			0.01	5	0.00	2	0.00	1	0.00	1	0.01	6	0.0198	15
Promoter	99278224	−705 3 base pair deletion			0.00	1	0.00	1	0.01	5	0.00	1	0.00	3	0.0146	11
Promoter	99278223	−704 A>G			0.00	0	0.00	0	0.00	0	0.00	1	0.00	0	0.0013	1
Promoter	99278152	−633 C>A			0.00	0	0.00	0	0.00	0	0.00	0	0.00	1	0.0013	1
Promoter	99278146	−627 G>A			0.00	0	0.00	0	0.00	1	0.00	0	0.00	0	0.0013	1
Promoter	99278144	−625 A>G			0.00	0	0.00	0	0.00	0	0.00	1	0.00	0	0.0013	1
Promoter	99278070	−551 C>A	rs28365079		0.01	4	0.01	5	0.02	15	0.01	8	0.01	4	0.0476	36
Promoter	99277988	−469 G>A			0.00	0	0.00	0	0.00	0	0.00	1	0.00	0	0.0013	1
UTR of exon 1	99277593	−74 C>T	rs28371764		0.00	2	0.01	6	0.00	0	0.00	2	0.00	2	0.0158	12
UTR of exon 1	99277544	−25 A>C			0.00	0	0.00	0	0.00	0	0.00	0	0.00	1	0.0013	1
Intron 1	99277392	127 G>A			0.00	0	0.00	0	0.00	1	0.00	2	0.00	0	0.0040	3
Intron 1	99277337	182 C>A			0.00	0	0.00	0	0.00	3	0.00	0	0.00	0	0.0040	3
Intron 2	99272310	5209 C>T	rs28365067		0.01	11	0.02	12	0.01	5	0.01	8	0.01	8	0.0580	44
Intron 2	99272290	5229 G>A	rs41301652		0.00	0	0.00	0	0.00	2	0.00	0	0.00	0	0.0026	2
Intron 2	99272275	5244 C>T			0.00	0	0.00	0	0.00	0	0.00	0	0.00	2	0.0026	2
Intron 3	99272103	5416 C>T			0.00	0	0.00	0	0.00	0	0.00	0	0.00	2	0.0026	2
Intron 3	99272009	5510 T>A	rs28969392		0.01	6	0.01	4	0.01	10	0.01	9	0.00	3	0.0422	32
Intron 3	99271928	5591 C>T	rs41301655		0.00	0	0.01	4	0.00	1	0.00	0	0.00	2	0.0092	7
Intron 3	99271853	5666 A>G	rs41301658		0.00	1	0.00	1	0.00	3	0.01	7	0.00	2	0.0185	14
Intron 3	99271808	5711 A>G	rs41258334		0.01	11	0.01	11	0.01	5	0.01	9	0.01	8	0.0580	44
Intron 3	99271778	5741 A>G			0.01	6	0.00	3	0.01	4	0.01	8	0.00	3	0.0317	24
Intron 3	99270539	6980 A>G	rs776746	Defines the variant *CYP3A5*3*	0.13	95	0.14	102	0.06	44	0.10	75	0.13	97	0.5581	413
Intron 3	99270504	7015 3 base pair deletion			0.00	0	0.00	0	0.00	0	0.00	1	0.00	0	0.0014	1
Intron 3	99270318	7201 C>T	rs8175345		0.00	0	0.00	1	0.01	9	0.00	0	0.00	1	0.0149	11
Exon 4	99270249	7270 G>A		G77S	0.00	0	0.00	0	0.00	0	0.00	1	0.00	0	0.0014	1
Intron 4	99270164	7355 C>T	rs28365074		0.00	0	0.00	0	0.00	1	0.00	0	0.00	2	0.0041	3
Intron 5	99264352	13167 T>C	rs68178885		0.00	3	0.00	2	0.00	1	0.00	3	0.00	1	0.0132	10
Intron 6	99264149	13370 G>A	rs41301670		0.00	0	0.00	0	0.00	0	0.00	0	0.00	2	0.0027	2
Exon 7	99262835	14684 G>A	rs10264272	Defines the variant *CYP3A5*6*	0.04	28	0.03	23	0.05	39	0.03	23	0.03	21	0.1763	134

Table 3 All polymorphic sites identified in an 8063bp *CYP3A5* region re-sequenced in five Ethiopian populations *(Continued)*

Exon 7	99262793	14726 A>G	rs2838372	Synonymous	0.00	1	0.00	0	0.00	0	0.00	0	0.00	0	0.0013	1
Intron 7	99262642	14877 A>G			0.00	1	0.01	5	0.02	12	0.01	9	0.00	2	0.0382	29
Intron 7	99261737	15782 T>C	rs28969393		0.01	5	0.01	4	0.01	9	0.01	9	0.00	3	0.0396	30
Exon 8	99261651	15868 A>G		K266R	0.00	0	0.00	0	0.00	1	0.00	0	0.00	0	0.0013	1
Intron 8	99261583	15936 C>A			0.00	0	0.00	0	0.00	0	0.00	2	0.00	0	0.0026	2
Intron 8	99260546	16973 G>A			0.00	0	0.00	1	0.00	0	0.00	0	0.00	0	0.0013	1
Exon 9	99260502	17017 C>T		R268Stop	0.00	0	0.00	0	0.00	1	0.00	0	0.00	0	0.0013	1
Intron 9	99260407	17112 C>T	rs28383478		0.00	0	0.00	2	0.00	0	0.00	0	0.00	0	0.0026	2
Intron 9	99260362	17157 G>T	rs4646453		0.00	3	0.00	3	0.01	4	0.01	10	0.01	5	0.0331	25
Intron 9	99260282	17237 T>G			0.00	0	0.00	0	0.00	1	0.00	0	0.00	0	0.0013	1
Intron 9	99260170	17349 T>G			0.00	3	0.00	2	0.01	7	0.01	7	0.00	3	0.0291	22
Intron 9	99258524	18995 C>T	rs10247580		0.00	0	0.00	2	0.02	12	0.01	7	0.00	1	0.0291	22
Intron 9	99258320	19199 G>A			0.00	0	0.00	0	0.00	1	0.00	0	0.00	0	0.0013	1
Intron 9	99258316	19203 T>C			0.00	0	0.00	0	0.00	0	0.00	2	0.00	0	0.0026	2
Exon 10	99258124	19395 A>C		K342T	0.00	0	0.00	0	0.00	0	0.00	0	0.00	1	0.0013	1
Exon 11	99250397	27125-27126 T insertion	rs41303343	Defines the variant *CYP3A5*7*	0.00	0	0.00	0	0.00	1	0.00	1	0.00	0	0.0026	2
Exon 11	99250381	27138 A>G		V350M	0.00	0	0.00	0	0.00	1	0.00	0	0.00	0	0.0013	1
Intron 12	99247647	29872 G>T			0.00	0	0.00	0	0.00	0	0.00	2	0.00	0	0.0026	2
Intron 12	99247503	30016 1 base pair deletion	rs28365093		0.00	3	0.01	4	0.02	15	0.01	8	0.01	4	0.0450	34
Intron 12	99246026	31493 T>C	rs28365069		0.01	4	0.01	11	0.01	11	0.02	18	0.01	9	0.0699	53
3' UTR	99245914	31605 C>T	rs15524		0.14	105	0.14	109	0.09	69	0.11	84	0.14	107	0.6253	474

n refers to the total number of chromosomes on which a particular variant was observed. *f* is the relative frequency of each variant. Total refers to the number of times a variant was observed in the Ethiopian cohort (758 chromosomes) and *f* is its relative frequency. Position on chromosome 7 is based on NCBI Build 132, February 2009.

more detailed surveys of clinically important variation in diverse African populations will improve our understanding of how specific drugs and dosages contribute to adverse clinical outcomes within Africa and the African Diaspora. The number of such studies will undoubtedly increase with the availability of newer and cheaper sequencing technologies [58,59] and progression towards the $1000 genome [60].

We combined our African *CYP3A5*3* data with those previously published to examine the global prevalence of the allele. We found a significant, positive correlation between *CYP3A5*3* allele frequencies and latitude, consistent with a previous report [20]. This correlation remained significant when only African data were considered [Spearmann Rho= 0.666, $p<0.0001$]. In contrast we found no significant correlation between latitude and *CYP3A5*6* or *CYP3A5*7* frequencies. Given the restricted geographic distribution of the *CYP3A5*6* allele mainly to Africa, coupled with our estimates of its age (>200,000 years), it is possible that this allele was lost in a population bottleneck during the range-expansion of humans out of Africa. The heterogeneous distribution of *CYP3A5*7* in Africa suggests that it arose from a much more recent mutation event and may have spread with the expansion of Niger-Congo speaking populations ~4000 years ago [61]. Nonetheless, the reasons why the derived *CYP3A5*3*, *CYP3A5*6* and *CYP3A5*7* alleles are found at appreciable frequencies in sub-Saharan Africa remains unknown, and the possibility of independent evolutionary causes cannot be discounted. The global distribution of the *CYP3A5*3* allele is unusual when compared with microsatellite markers, genotyped in samples from the Human Genome Diversity Panel (HDGP-CEPH) [20]. Integrated haplotype scores (iHS) for *CYP3A5*3* haplotypes in HGDP-CEPH populations sampled from high latitudes north and south of the equator are outliers in the iHS genome-wide distribution (iHS≥2) [62]. iHS scores of *CYP3A5*3* haplotypes are similar to those of genomic regions surrounding the *LCT* (lactase) and *CD36* genes [63], which have both been reported to have undergone positive selection [64-66]. It is plausible that an increase in latitude of ~20°, when humans first expanded from East Africa to the Arabian Peninsula, is coupled with specific environmental changes which provided a novel selection pressure. Temperature and precipitation data associated with the Quaternary QUEST project (accessed through the British Atmospheric Centre: http://badc.nerc.ac.uk/home/index.html) suggest that changes in precipitation over the past 50,000 years are greater than those in temperature. However we did not find any significant correlation between precipitation values and *CYP3A5* allele frequencies. We did observe negative correlations between inferred expression phenotypes (assuming *CYP3A5*6* is a low/non-expresser allele) and aridity values for the Holocene [Spearmann's Rho= −0.465, $p=0.004$] and Late Pleistocene [Spearmann's Rho= −0.379, $p=0.02$]. Under the de Martonne aridity index, this means that high frequencies of high-activity alleles are positively correlated with arid and semi-arid environments [34]. This finding is consistent with the hypothesis that high-activity *CYP3A5* alleles may be adaptive in regions where there are frequent water shortages, by aiding the rapid retention of water [20]. However, stronger correlations were found with temperature alone. Although further work will be needed to confirm these ecological correlations, the strong correlation with temperature is consistent with what we would expect for functional variation of genes involved in heat adaptation [21]. However, we cannot rule out that there may be an, as yet untested, ecological variable which may have provided a selective pressure.

We have provided the first estimate of the age of the *CYP3A5*3* allele as ~76,000 years (95% C.I. 57,504-

Table 4 A summary of the tests for departures from neutrality for an 8063bp region of *CYP3A5*

	Global populations			Ethiopian populations				
	African-Americans	Europeans	Han Chinese	Afar	Amhara	Anuak	Maale	Oromo
Sample size	23	24	23	75	76	76	76	76
Nucleotide diversity (π)	5.4×10^{-4}	9×10^{-5}	1.1×10^{-4}	2.1×10^{-4}	2.5×10^{-4}	3.6×10^{-4}	3.5×10^{-4}	2.2×10^{-4}
McDonald-Kreitman test	0.475	0.50	0.777	0.462	0.475	0.576	1.00	0.777
Tajima's *D*	−1.04	−1.92	−1.21	−1.46	−1.13	−1.26	−0.96	−1.79
Fu and Li's *D**	−0.97	**−2.86**	−1.82	−1.19	0.12	−2.72	−0.11	−1.05
Fu and Li's *F**	−1.17	**−3.00**	−1.91	−1.54	−0.43	−2.57	−0.53	−1.58
Fu and Li's *D*	−0.64	−1.37	−1.45	−1.56	−0.08	−1.93	0.73	−0.96
Fu and Li's *F*	−0.92	−1.81	−1.55	−1.79	−0.52	−1.96	0.13	−1.48
Fu's F_S	**−31.06**	**−5.71**	−1.54	**−9.48**	**−11.74**	**−18.44**	**−11.08**	**−22.84**
Strobeck's *S*	**1.00**	**0.999**	0.929	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
Fay and Wu's *H* statistic	−0.13140	−3.25532	−1.89372	0.10774	−0.23998	−0.47177	−0.87086	−1.75514

Statistically significant departures from neutrality, following Bonferonni correction (correction for 8 tests; adjusted $p \leq 0.00625$) are shown in bold.

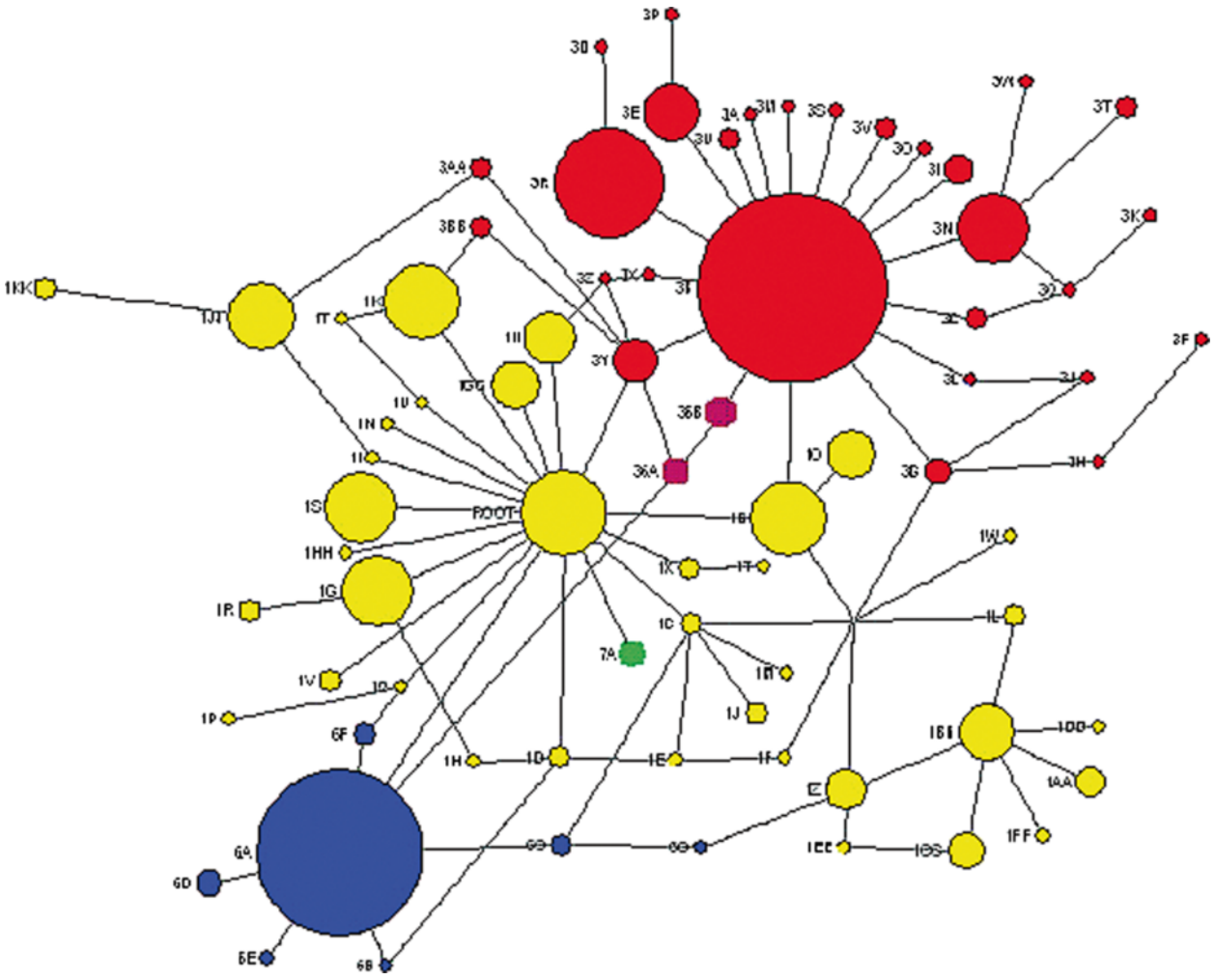


Figure 2 A network of all *CYP3A5* haplotypes inferred for Ethiopians and Coriell populations. Networks assume single mutational steps. Haplotypes are colored according to the haplogroup to which they belong; haplotypes which are defined by the *CYP3A5*3* allele are shown in red; those defined by *CYP3A5*1* in yellow; *CYP3A5*7* in green; *CYP3A5*6* in blue and *CYP3A5*3/*6* recombinant haplotypes are shown in purple. The size of each haplotype is proportional to its frequency in the global database. Additional file 4 Figure S3a contains information on the exact composition of each coded haplotype.

102,752 years), using a closely linked microsatellite. This estimate is consistent with the wide geographic distribution of *CYP3A5*3*, both inside and outside Africa although not with the low haplotype diversity [20,67] and high iHS scores [62] previously reported. This may be because the allele age estimation presented here was made using Ethiopian data whereas the previously reported low haplotype diversity [20,67] and high iHS scores [62] were based on non-African samples. It is also possible that the microsatellite mutation rate for rs10536492 differs from the genome wide average for dinucleotide repeats [68], which would influence the age estimates, or that the extensive LD on haplotypes containing *CYP3A5*3* is explained by unusually low recombination rates in that genomic region. Finally, while the correlations between the *CYP3A5*3* geographic distribution and ecological variables relating to temperature and aridity, as well as previously reported low haplotype diversity [20,67] and high iHS scores [62], both independently support the hypothesis that the *CYP3A5*3* allele is adaptive, alternative targets of positive selection on the same haplotype background may exist and both they (if present) and the *CYP3A5*3* allele may have been differentially selected inside and outside Africa.

Conclusions

The data that we present complement and extend work in previous publications which reported evidence of geographically restricted positive selection on the *CYP3A5* gene [20,25,26,69]. In addition to improved knowledge of the effect and distribution of clinically relevant genetic variation, our approach highlights the importance of

considering population history and of utilizing evolutionary approaches in clinical research. Evolutionary approaches to genetic studies are likely to identify additional populations that require targeted health interventions. Further studies which characterize variation in medically important genes in ethnically and geographically diverse global populations are needed as we progress towards personalized clinical medicine, a key goal of the genomics revolution [70].

Methods

Samples

The DNA samples analyzed in this study were part of a collection at The Centre for Genetic Anthropology at University College London. Samples were collected anonymously and with informed consent (verbal in Africa) from ostensibly healthy individuals, between 1998–2007, from specified locations in and around Africa [ethical approval: UCLH 99/0196]. Additional ethical approval was obtained for Ethiopian collections from the National Health Research Ethical Clearance Committee under the Ethiopian Science and Technology Commission in Addis Ababa. All samples have been previously used in studies on clinically relevant genes [31-33]. For analyses, individuals were grouped by the collection location or by ethnicity (Additional file 5 Table S2). Samples were not grouped according to country as the partitioning of much of the African continent by colonial powers was recent and largely irrespective of ethnic identities [71]. 1028 *CYP3A5*1/*3* genotypes for 51 global populations, from the Human Genome Diversity Panel-Centre d'Etude du Polymorphisme Humain (HGDP-CEPH) collection, which had previously been published [20] were combined with the 2538 sample cohort genotyped for this study. *CYP3A5* re-sequencing data, which were previously published, for 70 individuals from three distinct ethnic groups from the Coriell Repositories (24 European Caucasians, 23 African-Americans and 23 Han Chinese individuals) were combined with the Ethiopian cohort for detailed integrative analyses [20]. Published data were provided by Dr Emma Thompson from the University of Chicago.

Genotyping and re-sequencing

Genotyping of clinically relevant CYP3A5 alleles; Genotyping of *CYP3A5*1, CYP3A5*3, CYP3A5*6* and *CYP3A5*7* was performed using TaqMAN allelic discrimination technology [ABiosystems product code: C_26201809_30 for *CYP3A5*1/*3,* and ABiosystems product code: C_30203959 for *CYP3A5*6*], and KASPar (performed externally by KBiosciences®, UK).

Re-sequencing of CYP3A5; The 13 exons and their flanking introns, promoter region and 3′ untranslated of *CYP3A5* were amplified in 379 Ethiopian individuals using primers designed on the basis of the *CYP3A5* reference sequence in NCBI Build 132 [(http://www.ncbi.nlm.nih.gov/) (see Additional file 6 Table S3 for a list of primers)]. Amplicons were sequenced using ABI PRISM Dye Terminators version 3.1 on an ABI 96-capillary 3730×l DNA Analyzer according to the manufacturer's protocol (Applied Biosystems, Applera, UK). Part of the *CYP3A5* gene was re-sequenced externally by Macrogen®, USA.

Microsatellite genotyping; A –GT microsatellite, located ~1500 base pairs downstream of the 3′ end of *CYP3A5* was genotyped in 379 Ethiopian individuals, for whom re-sequencing data were also generated. Microsatellite genotyping was performed using a high-throughput method adapted from [72]. A 456 base pair region of *CYP3A5,* approximately ~1000 base pairs downstream of the 3′ UTR was amplified using the forward primer 5′-AATATATGTGTTTGTATGTGTG-3′ and a fluorescently labeled reverse primer FAM-AAGTGCTACCAATTTTGT ACGT-3′. PCR amplification was performed in 10 μl reaction volumes containing 1ng of template DNA, 0.5 μM of primers, 0.2 units *Taq* DNA polymerase (HT Biotech, Cambridge, UK), 0.2 μmol dNTPs, 0.1 μmol of 10X Buffer IV (Thermo Scientific®) and 0.28 μl of magnesium chloride (concentration 25 mM). Cycling conditions were 5 minutes of pre-incubation at 95°C, followed by 38 cycles of 95°C for one minute, 58°C for 40 seconds, 72°C for 40 seconds, with

Table 5 Pairwise F_{ST} values for five Ethiopian populations and three other global populations

	Afar	Amhara	Anuak	Maale	Oromo	African-Americans	Europeans	Han Chinese
Afar	*	0.74597	**<0.00001**	**0.00436**	0.76537	**<0.00001**	**<0.00001**	**<0.00001**
Amhara	−0.00248	*	**<0.00001**	**0.00347**	0.93525	**<0.00001**	**<0.00001**	**<0.00001**
Anuak	0.04566	0.05138	*	**0.00257**	**<0.00001**	**<0.00001**	**<0.00001**	**<0.00001**
Maale	0.01951	0.01736	0.01061	*	**0.00267**	**<0.00001**	**<0.00001**	**<0.00001**
Oromo	−0.00255	−0.0036	0.04981	0.01547	*	**<0.00001**	**<0.00001**	**<0.00001**
African-Americans	0.08997	0.09366	0.01558	0.03432	0.08803	*	**<0.00001**	**<0.00001**
Europeans	0.04873	0.03807	0.15448	0.08716	0.03371	0.19028	*	**0.00257**
Han Chinese	0.10812	0.0893	0.23763	0.16215	0.09715	0.29154	0.0677	*

Pairwise F_{ST} values are shown in the bottom left side of the Table, the corresponding *p*-values are shown in the top right of the Table. *P*-values which are significant after Bonferroni correction (adjusted *p*-value = 0.00625; correction for 8 tests) are shown in bold.

Table 6 Age estimates of clinically relevant *CYP3A5* variants

CYP3A5 variant	Location on chromosome 7	Location in gene	Allele dated	Number of chromosomes	Average squared distance (ASD)	Time to most recent common ancestor		95% confidence intervals of allele age estimate based on a star phylogeny			
						Estimate of allele age		Lower		Upper	
						Generations	Years	Generations	Years	Generations	Years
CYP3A5*3	99270539	Intron 3	G	134	1.0746	2388	76,416	1797	57,504	3211	102,752
CYP3A5*6	99262835	Exon 7	A	18	3.0714	6825	218,400	3086	98.752	11975	383,200
rs15524	99245914	3′ UTR	T	324	1.8426	4095	131,040	3157	101,024	5413	173,216

Allele ages were estimated using a mutation rate of 0.00045 and a generation time of 32 years. The confidence intervals for the estimated age of the *CYP3A5*6* are large; most likely a reflection of the small sample size. UTR is the Untranslated Region.

a final elongation step at 72°C for 10 minutes. Following amplification, a 1.1 μl aliquot of amplified PCR product was added to 9.89 μl of high purity (HiDi) formamide and 0.11 μl of ROX-500 size standard (Applied Biosystems, Warrington, UK). Samples were run on an ×3730 DNA Analyzer and analyzed using GeneMapper 4 software (Applied Biosystems, Warrington UK).

Data analyses

Molecular diversity and Population genetics; exact tests of deviation from Hardy-Weinberg equilibrium (using 10,000 steps in a Markov chain), pairwise F_{ST}, and AMOVA, were all performed using Arlequin 3.5 [73]. Pairwise F_{ST} estimates were used to perform principal co-ordinates analysis in the R-programming environment using routines in the APE package. The D' measure of linkage disequilibrium was calculated using the expectation maximization algorithm using LDMax (part of the GOLD software package, freely available at: http://www.sph.umich.edu/csg/abecasis/GOLD/docs/ldmax.html).
Haplotypes were inferred using PHASE version 2.1 (1000 iterations, 500 burn-in) [74]. Singletons were removed for haplotype and LD analysis. Haplotype networks were constructed using a median-joining network implemented in Network 4.6.1 and re-colored using Adobe PhotoShop CS4. Nucleotide diversity, tests for departures from neutrality, Fay and Wu's H test and the HKA test were all performed using DnaSP 5.0 [75]. The chimpanzee *CYP3A5* gene sequence was downloaded from NCBI (http://www.ncbi.nlm.nih.gov/).

Ecological correlations; geographic co-ordinates were used to calculate distance from the equator (in kilometers) using the online programme: http://www.movable-type.co.uk/scripts/latlong.html. Raw ecological data for temperature (in degrees Celsius) and precipitation (in mm), at each set of geographic co-ordinates, for 0, 10,000 and 50,000 years ago were extracted from the British Atmospheric Data Centre (http://badc.nerc.ac.uk/home/index.html), from the ALL-5G dataset associated with the Quaternary QUEST (http://researchpages.net/QQ/) [76]. The data were extracted using Python. The raw data have a resolution of 5 degrees latitude and 7.5 degrees longitude and were interpolated to a resolution of 1 degree latitude and 1 degree longitude. The interpolations were done using the smooth.2d function in the fields library of the R-programming environment. An estimate of relative aridity was inferred from extracted temperature and precipitation values corresponding to each geographic location using the de Martonne aridity index [34]. Mantel and partial Mantel tests were performed in the R-programming environment using routines in the APE package [77] and ecodist package [78] respectively.

Bioinformatics analyses of genetic variation on protein expression and function; cross-species alignments of *CYP3A5* orthologues (sequences obtained from NCBI: http://www.ncbi.nlm.nih.gov/) were performed using ClustalW software (http://www.ebi.ac.uk/Tools/msa/clustalw2/). Analyses of regulatory motifs in the *CYP3A5* promoter were performed using MatInspector [79], effects of amino acid substitutions on the structure and function of CYP3A5 were performed using PolyPhen2 [80], predictions of mutations which are likely to affect gene splicing were performed using the online Berkeley Drosophila Genome Project splice predictor [81].

Estimating the age of the clinically relevant CYP3A5 variants; The gametic phase of CYP3A5 mutations and the –GT microsatellite (rs10536492) was not determined empirically. Allele ages were estimated using data for individuals homozygous for particular haplotypes. As no Ethiopian individual was identified to be a *CYP3A5*7* homozygote, this variant could not be dated. Under the stepwise mutation model the variance (ASD) in the microsatellite repeat length, from the most recent common ancestor, is a linear function of the mutation rate (μ) and coalescence time in generations (t); ASD=μt [82,83]. A mutation rate of 4.5×10^{-4} was used to estimate the time to the most recent common ancestor (MRCA)

based on average estimates of the mutation rate of dinucleotide microsatellites in the human genome [68,84].

ASD and t were calculated using Ytime software [85]. The microsatellite length of the ancestral MRCA is assumed to be known. For this study the ancestral length of the microsatellite was estimated to be 35; as the majority of *CYP3A5*1* haplotypes had 35 repeats. Confidence intervals for the age estimates were obtained from calculating the distances between the ancestral and derived chromosomes under a star-genealogy model; based on the results of network analysis of *CYP3A5* haplotypes. A generation was assumed to be 32 years [46].

Additional files

Additional file 1: Table S1."The proportion of each inferred *CYP3A5* haplotype observed in each population." The Table lists the frequencies of each inferred *CYP3A5* haplotype, by population.

Additional file 2: Figure S1. "The distribution of high-, intermediate- and low- CYP3A5 expression phenotypes, inferred from diplotypes." The Figure shows inferred CYP3A5 expression phenotypes, assuming that *CYP3A5*6* causes low/non-expression of CYP3A5. The size of each circle is proportional to the number of individuals sampled from a given population (see Additional file Table S1).

Additional file 3: Figure S2. "The distribution of high-, intermediate- and low- CYP3A5 expression phenotypes, inferred from diplotypes." The Figure shows inferred CYP3A5 expression phenotypes, assuming that *CYP3A5*6* does not cause low/non-expression of CYP3A5. The size of each circle is proportional to the number of individuals sampled from a given population (see Additional file Table S1).

Additional file 4: Figures S3a and b. "Haplotypes inferred from genotype data in 8 populations." Supplementary Figure 3a shows the composition of each *CYP3A5* haplotype inferred from genotype data for 8 global populations. The frequencies of each haplotype, by population, are shown in Additional file Figure S3b.

Additional file 5: Table S2. "Geographic co-ordinates, sample size and major language family of each population genotyped in the geographic survey of clinically relevant *CYP3A5* alleles. The *CYP3A5* gene was re-sequenced in five Ethiopian populations." This Table provides details of all populations which were genotyped, and re-sequenced for this study.

Additional file 6: Table S3. "A list of the primers used for PCR amplification and sequencing of *CYP3A5*."

Competing interest

Neil Bradman is Chairman of The Henry Stewart Group and London and City Group of Companies and has extensive business and financial interests including involvement in biotechnology ventures and educational material used by researchers in the life sciences. The research has been funded in part by the London and City Group of Companies and the Melford Charitable Trust of which Neil Bradman is a trustee. The Melford Charitable Trust, London and City Group of Companies and Neil Bradman do have any intellectual, or other, property rights whatsoever with respect to the research which forms the subject matter of the paper. All other authors have no conflict of interest.

Authors' contributions

RKB carried out the molecular genetic studies, analyses of data and drafted the manuscript. MK extracted and interpolated all climate data from the British Atmospheric Survey. CAP performed genotyping of the Hypervariable Segment 1 and the Y-chromosome in Ethiopian populations. AT and EB collected all Ethiopian samples which were used for analysis within this study. NNB assisted with the collection of most African samples used in this study, and conceived the initial experimental design of the project. MGT conceived the statistical analyses of the project, in particular those relating to ecological data, and oversaw the writing of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors wish to thank Professor Andrés Ruiz-Linares for help and guidance with the analyses, as well as Professor Dallas Swallow, Mr. Victor Acuña-Alonso, and Mr. Pawel Zmarz for helpful discussion of the manuscript, and Dr Emma Thompson for providing *CYP3A5*3* genotypes and *CYP3A5* re-sequencing data for combined analyses. Additional thanks to Ranji Arasaretnam, Mari-Wyn Burley and Rosemary Ekong for help with DNA extractions and sequencing. This work was supported by a Biotechnology and Biological Sciences Research Council-CASE Ph. D. studentship, funding from the London and City Group of Companies and the Melford Charitable Trust [which were awarded to Ripudaman K Bains], and funding from the Engineering and Physical Sciences Research Council [which was awarded to Mirna Kovacevic through UCL CoMPLEX].

Author details

[1]Research Department of Genetics, Evolution and Environment, University College London, Darwin Building, Gower Street, London WC1E 6BT, UK. [2]Centre for Mathematics and Physics in the Life Sciences and Experimental Biology (CoMPLEX), University College London, Physics Building, Gower Street, London WC1E 6BT, UK. [3]Addis Ababa University and Center of Human Genetic Diversity, P.O. Box 1176, Addis Ababa, Ethiopia. [4]Henry Stewart Group, 28/30 Little Russell Street, London WC1A 2HN, UK. [5]Department of Evolutionary Biology, Uppsala University, Uppsala, Sweden.

Received: 15 February 2013 Accepted: 25 April 2013
Published: 3 May 2013

References

1. Coleman R: **Disease burden in sub-Saharan Africa.** *Lancet* 1998, **351**(9110):1208.
2. Stearns SC: **Evolutionary medicine: its scope, interest and potential.** *Proceedings Biological sciences/The Royal Society* 2012, **279**(1746):4305–4321.
3. Zhou SF, Liu JP, Chowbay B: **Polymorphism of human cytochrome P450 enzymes and its clinical impact.** *Drug Metab Rev* 2009, **41**(2):89–295.
4. Nebert DW, Russell DW: **Clinical importance of the cytochromes P450.** *Lancet* 2002, **360**(9340):1155–1162.
5. Wojnowski L: **Genetics of the variable expression of CYP3A in humans.** *Ther Drug Monit* 2004, **26**(2):192–199.
6. Perera MA: **The missing linkage: what pharmacogenetic associations are left to find in CYP3A?** *Expert Opin Drug Metab Toxicol* 2010, **6**(1):17–28.
7. Shi Y, Li Y, Tang J, Zhang J, Zou Y, Cai B, Wang L: **Influence of CYP3A4, CYP3A5 and MDR-1 polymorphisms on tacrolimus pharmacokinetics and early renal dysfunction in liver transplant recipients.** *Gene* 2013, **512**(2):226–231.
8. Bochud M, Eap CB, Elston RC, Bovet P, Maillard M, Schild L, Shamlaye C, Burnier M: **Association of CYP3A5 genotypes with blood pressure and renal function in African families.** *J Hypertens* 2006, **24**(5):923–929.
9. Plummer SJ, Conti DV, Paris PL, Curran AP, Casey G, Witte JS: **CYP3A4 and CYP3A5 genotypes, haplotypes, and risk of prostate cancer.** *Canc Epidemiol Biomarkers Prev: a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* 2003, **12**(9):928–932.
10. Lee SJ, Usmani KA, Chanas B, Ghanayem B, Xi T, Hodgson E, Mohrenweiser HW, Goldstein JA: **Genetic findings and functional studies of human CYP3A5 single nucleotide polymorphisms in different ethnic groups.** *Pharmacogenetics* 2003, **13**(8):461–472.
11. Aoyama T, Yamano S, Waxman DJ, Lapenson DP, Meyer UA, Fischer V, Tyndale R, Inaba T, Kalow W, Gelboin HV: **Cytochrome P-450 hPCN3, a novel cytochrome P-450 IIIA gene product that is differentially expressed in adult human liver. cDNA and deduced amino acid sequence and distinct specificities of cDNA-expressed hPCN1 and hPCN3 for the metabolism of steroid hormones and cyclosporine.** *J Biol Chem* 1989, **264**(18):10388–10395.
12. Schuetz JD, Molowa DT, Guzelian PS: **Characterization of a cDNA encoding a new member of the glucocorticoid-responsive cytochromes P450 in human liver.** *Arch Biochem Biophys* 1989, **274**(2):355–365.
13. Kuehl P, Zhang J, Lin Y, Lamba J, Assem M, Schuetz J, Watkins PB, Daly A, Wrighton SA, Hall SD, Maurel P, Relling M, Brimer C, Yasuda K, Venkataramanan R, Strom S, Thummel K, Boguski MS, Schuetz E: **Sequence**

diversity in CYP3A promoters and characterization of the genetic basis of polymorphic CYP3A5 expression. *Nat Genet* 2001, **27**(4):383–391.
14. Hustert E, Haberl M, Burk O, Wolbold R, He YQ, Klein K, Nuessler AC, Neuhaus P, Klattig J, Eiselt R, Koch I, Zibat A, Brockmoller J, Halpert JR, Zanger UM, Wojnowski L: **The genetic determinants of the CYP3A5 polymorphism.** *Pharmacogenetics* 2001, **11**(9):773–779.
15. Quaranta S, Chevalier D, Allorge D, Lo-Guidice JM, Migot-Nabias F, Kenani A, Imbenotte M, Broly F, Lacarelle B, Lhermitte M: **Ethnic differences in the distribution of CYP3A5 gene polymorphisms.** *Xenobiotica* 2006, **36**(12):1191–1200.
16. Roy JN, Lajoie J, Zijenah LS, Barama A, Poirier C, Ward BJ, Roger M: **CYP3A5 genetic polymorphisms in different ethnic populations.** *Drug Metab Dispos* 2005, **33**(7):884–887.
17. Park SY, Kang YS, Jeong MS, Yoon HK, Han KO: **Frequencies of CYP3A5 genotypes and haplotypes in a Korean population.** *J Clin Pharm Ther* 2008, **33**(1):61–65.
18. He P, Court MH, Greenblatt DJ, Von Moltke LL: **Genotype-phenotype associations of cytochrome P450 3A4 and 3A5 polymorphism with midazolam clearance *in vivo*.** *Clin Pharmacol Ther* 2005, **77**(5):373–387.
19. Tucker AN, Tkaczuk KA, Lewis LM, Tomic D, Lim CK, Flaws JA: **Polymorphisms in cytochrome P4503A5 (CYP3A5) may be associated with race and tumor characteristics, but not metabolism and side effects of tamoxifen in breast cancer patients.** *Cancer Lett* 2005, **217**(1):61–72.
20. Thompson EE, Kuttab-Boulos H, Witonsky D, Yang L, Roe BA, Di Rienzo A: **CYP3A variation and the evolution of salt-sensitivity variants.** *Am J Hum Genet* 2004, **75**(6):1059–1069.
21. Young JH, Chang YP, Kim JD, Chretien JP, Klag MJ, Levine MA, Ruff CB, Wang NY, Chakravarti A: **Differential susceptibility to hypertension is due to selection during the out-of-Africa expansion.** *PLoS Genet* 2005, **1**(6):e82.
22. Wrighton SA, Brian WR, Sari MA, Iwasaki M, Guengerich FP, Raucy JL, Molowa DT, Vandenbranden M: **Studies on the expression and metabolic capabilities of human liver cytochrome P450IIIA5 (HLp3).** *Mol Pharmacol* 1990, **38**(2):207–213.
23. Bochud M, Bovet P, Burnier M, Eap CB: **CYP3A5 and ABCB1 genes and hypertension.** *Pharmacogenomics* 2009, **10**(3):477–487.
24. Givens RC, Lin YS, Dowling AL, Thummel KE, Lamba JK, Schuetz EG, Stewart PW, Watkins PB: **CYP3A5 genotype predicts renal CYP3A activity and blood pressure in healthy adults.** *J Appl Physiol* 2003, **95**(3):1297–1300.
25. Chen X, Wang H, Zhou G, Zhang X, Dong X, Zhi L, Jin L, He F: **Molecular population genetics of human CYP3A locus: signatures of positive selection and implications for evolutionary environmental medicine.** *Environ Health Perspect* 2009, **117**(10):1541–1548.
26. Li J, Zhang L, Zhou H, Stoneking M, Tang K: **Global patterns of genetic diversity and signals of natural selection for human ADME genes.** *Hum Mol Genet* 2011, **20**(3):528–540.
27. DeGiorgio M, Jakobsson M, Rosenberg NA: **Out of Africa: modern human origins special feature: explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from Africa.** *Proc Natl Acad Sci USA* 2009, **106**(38):16057–16062.
28. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, Myers RM: **Worldwide human relationships inferred from genome-wide patterns of variation.** *Science* 2008, **319**(5866):1100–1104.
29. Prugnolle F, Manica A, Balloux F: **Geography predicts neutral genetic diversity of human populations.** *Curr Biol* 2005, **15**(5):R159–R160.
30. Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL: **Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa.** *Proc Natl Acad Sci USA* 2005, **102**(44):15942–15947.
31. Browning SL, Tarekegn A, Bekele E, Bradman N, Thomas MG: **CYP1A2 is more variable than previously thought: a genomic biography of the gene behind the human drug-metabolizing enzyme.** *Pharmacogenet Genomics* 2010, **20**(11):647–664.
32. Horsfall LJ, Zeitlyn D, Tarekegn A, Bekele E, Thomas MG, Bradman N, Swallow DM: **Prevalence of clinically relevant UGT1A alleles and haplotypes in African populations.** *Ann Hum Genet* 2011, **75**(2):236–246.
33. Veeramah KR, Thomas MG, Weale ME, Zeitlyn D, Tarekegn A, Bekele E, Mendell NR, Shephard EA, Bradman N, Phillips IR: **The potentially deleterious functional variant flavin-containing monooxygenase 2*1 is at high frequency throughout sub-Saharan Africa.** *Pharmacogenet Genomics* 2008, **18**(10):877–886.
34. Croitoru A-E, Piticar A, Imbroane AM, Burada DC: **Spatiotemporal distribution of aridity indices based on temperature and precipitation in the extra-Carpathian regions of Romania.** *Theor Appl Climatol* 2012.
35. Nei M, Kumar S: *Molecular evolution and phylogenetics.* New York: Oxford University Press; 2000.
36. Kitano T, Liu YH, Ueda S, Saitou N: **Human-specific amino acid changes found in 103 protein-coding genes.** *Mol Biol Evol* 2004, **21**(5):936–944.
37. Burk O, Koch I, Raucy J, Hustert E, Eichelbaum M, Brockmoller J, Zanger UM, Wojnowski L: **The induction of cytochrome P450 3A5 (CYP3A5) in the human liver and intestine is mediated by the xenobiotic sensors pregnane X receptor (PXR) and constitutively activated receptor (CAR).** *J Biol Chem* 2004, **279**(37):38379–38385.
38. Lin YS, Dowling AL, Quigley SD, Farin FM, Zhang J, Lamba J, Schuetz EG, Thummel KE: **Co-regulation of CYP3A4 and CYP3A5 and contribution to hepatic and intestinal midazolam metabolism.** *Mol Pharmacol* 2002, **62**(1):162–172.
39. Busi F, Cresteil T: **CYP3A5 mRNA degradation by nonsense-mediated mRNA decay.** *Mol Pharmacol* 2005, **68**(3):808–815.
40. Kimura M: **The neutral theory of molecular evolution.** *Sci Am* 1979, **241**(5):98–100. 102, 108 passim.
41. Fay JC, Wu CI: **Hitchhiking under positive darwinian selection.** *Genetics* 2000, **155**(3):1405–1413.
42. Plaster CA: *Variation in Y chromosome, mitochondrial DNA, and labels of identity in Ethiopia.* University College London, Research Department of Genetics, Evolution and Environment: PhD thesis; 2011.
43. Pagani L, Kivisild T, Tarekegn A, Ekong R, Plaster C, Gallego Romero I, Ayub Q, Mehdi SQ, Thomas MG, Luiselli D, Bekele E, Bradman N, Balding DJ, Tyler-Smith C: **Ethiopian genetic diversity reveals linguistic stratification and complex influences on the Ethiopian gene pool.** *Am J Hum Genet* 2012, **91**(1):83–96.
44. Slatkin M, Rannala B: **Estimating allele age.** *Annu Rev Genomics Hum Genet* 2000, **1**:225–249.
45. Rannala B, Bertorelle G: **Using linked markers to infer the age of a mutation.** *Hum Mutat* 2001, **18**(2):87–100.
46. Tremblay M, Vezina H: **New estimates of intergenerational time intervals for the calculation of age and origins of mutations.** *Am J Hum Genet* 2000, **66**(2):651–658.
47. Griffiths RC, Marjoram P: **Ancestral inference from samples of DNA sequences with recombination.** *J Comput Biol* 1996, **3**(4):479–502.
48. Cruciani F, Santolamazza P, Shen P, Macaulay V, Moral P, Olckers A, Modiano D, Holmes S, Destro-Bisol G, Coia V, Wallace DC, Oefner PJ, Torroni A, Cavalli-Sforza LL, Scozzari R, Underhill PA: **A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes.** *Am J Hum Genet* 2002, **70**(5):1197–1214.
49. Salas A, Richards M, De la Fe T, Lareu MV, Sobrino B, Sanchez-Diz P, Macaulay V, Carracedo A: **The making of the African mtDNA landscape.** *Am J Hum Genet* 2002, **71**(5):1082–1111.
50. Lovell A, Moreau C, Yotova V, Xiao F, Bourgeois S, Gehl D, Bertranpetit J, Schurr E, Labuda D: **Ethiopia: between Sub-Saharan Africa and western Eurasia.** *Ann Hum Genet* 2005, **69**(Pt 3):275–287.
51. Gebeyehu E, Engidawork E, Bijnsdorp A, Aminy A, Diczfalusy U, Aklillu E: **Sex and CYP3A5 genotype influence total CYP3A activity: high CYP3A activity and a unique distribution of CYP3A5 variant alleles in Ethiopians.** *Pharmacogenomics J* 2011, **11**(2):130–137.
52. Kivisild T, Reidla M, Metspalu E, Rosa A, Brehm A, Pennarun E, Parik J, Geberhiwot T, Usanga E, Villems R: **Ethiopian mitochondrial DNA heritage: tracking gene flow across and around the gate of tears.** *Am J Hum Genet* 2004, **75**(5):752–770.
53. Mukonzo JK, Waako P, Ogwal-Okeng J, Gustafsson LL, Aklillu E: **Genetic variations in ABCB1 and CYP3A5 as well as sex influence quinine disposition among Ugandans.** *Ther Drug Monit* 2010, **32**(3):346–352.
54. Josephson F, Allqvist A, Janabi M, Sayi J, Aklillu E, Jande M, Mahindi M, Burhenne J, Bottiger Y, Gustafsson LL, Haefeli WE, Bertilsson L: **CYP3A5 genotype has an impact on the metabolism of the HIV protease inhibitor saquinavir.** *Clin Pharmacol Ther* 2007, **81**(5):708–712.
55. Oliveira E, Pereira R, Amorim A, McLeod H, Prata MJ: **Patterns of pharmacogenetic diversity in African populations: role of ancient and recent history.** *Pharmacogenomics* 2009, **10**(9):1413–1422.

56. Sgaier SK, Jha P, Mony P, Kurpad A, Lakshmi V, Kumar R, Ganguly NK: **Public health. Biobanks in developing countries: needs and feasibility.** *Science* 2007, **318**(5853):1074–1075.
57. Lamba JK, Lin YS, Schuetz EG, Thummel KE: **Genetic contribution to variable human CYP3A-mediated metabolism.** *Adv Drug Deliv Rev* 2002, **54**(10):1271–1294.
58. Butler D: **Genomics. Are you ready for the revolution?** *Nature* 2001, **409**(6822):758–760.
59. Chung RT: **Reaping the early harvest of the genomics revolution.** *Gastroenterology* 2010, **138**(5):1653–1654.
60. Defrancesco L: **Life Technologies promises $1,000 genome.** *Nat Biotechnol* 2012, **30**(2):126.
61. Beleza S, Gusmao L, Amorim A, Carracedo A, Salas A: **The genetic legacy of western Bantu migrations.** *Hum Genet* 2005, **117**(4):366–375.
62. Voight BF, Kudaravalli S, Wen X, Pritchard JK: **A map of recent positive selection in the human genome.** *PLoS Biol* 2006, **4**(3):e72.
63. Conrad DF, Jakobsson M, Coop G, Wen X, Wall JD, Rosenberg NA, Pritchard JK: **A worldwide survey of haplotype variation and linkage disequilibrium in the human genome.** *Nat Genet* 2006, **38**(11):1251–1260.
64. Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, Lander ES: **Positive natural selection in the human lineage.** *Science* 2006, **312**(5780):1614–1620.
65. Gerbault P, Liebert A, Itan Y, Powell A, Currat M, Burger J, Swallow DM, Thomas MG: **Evolution of lactase persistence: an example of human niche construction.** *Philos Trans R Soc Lond B Biol Sci* 2011, **366**(1566):863–877.
66. Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Powell K, Mortensen HM, Hirbo JB, Osman M, Ibrahim M, Omar SA, Lema G, Nyambo TB, Ghori J, Bumpstead S, Pritchard JK, Wray GA, Deloukas P: **Convergent adaptation of human lactase persistence in Africa and Europe.** *Nat Genet* 2007, **39**(1):31–40.
67. Thompson EE, Kuttab-Boulos H, Yang L, Roe BA, Di Rienzo A: **Sequence diversity and haplotype structure at the human CYP3A cluster.** *Pharmacogenomics J* 2006, **6**(2):105–114.
68. Whittaker JC, Harbord RM, Boxall N, Mackay I, Dawson G, Sibly RM: **Likelihood-based estimation of microsatellite mutation rates.** *Genetics* 2003, **164**(2):781–787.
69. Biswas S, Akey JM: **Genomic insights into positive selection.** *Trends in genetics: TIG* 2006, **22**(8):437–446.
70. Samani NJ, Tomaszewski M, Schunkert H: **The personal genome–the future of personalised medicine?** *Lancet* 2010, **375**(9725):1497–1498.
71. Pakenham T: *The scramble for Africa, 1876–1912.* London: Weidenfeld and Nicolson; 1991.
72. Thomas MG, Bradman N, Flinn HM: **High throughput analysis of 10 microsatellite and 11 diallelic polymorphisms on the human Y-chromosome.** *Hum Genet* 1999, **105**(6):577–581.
73. Excoffier L, Laval G, Schneider S: **Arlequin (version 3.0): an integrated software package for population genetics data analysis.** *Evol Bioinform Online* 2005, **1**:47–50.
74. Stephens M, Smith NJ, Donnelly P: **A new statistical method for haplotype reconstruction from population data.** *Am J Hum Genet* 2001, **68**(4):978–989.
75. Librado P, Rozas J: **DnaSP v5: a software for comprehensive analysis of DNA polymorphism data.** *Bioinformatics* 2009, **25**(11):1451–1452.
76. Smith RS, Gregory J: **The last glacial cycle: transient simulations with an AOGCM.** *Clim Dyn* 2012, **38:**1545–1560.
77. Paradis E, Claude J, Strimmer K: **APE: analyses of phylogenetics and evolution in R language.** *Bioinformatics* 2004, **20**(2):289–290.
78. Urban SCGDL: **The ecodist package for dissimilarity-based analysis of ecological data.** *J Stat Software* 2007, **22**(7):19.
79. Cartharius K, Frech K, Grote K, Klocke B, Haltmeier M, Klingenhoff A, Frisch M, Bayerlein M, Werner T: **MatInspector and beyond: promoter analysis based on transcription factor binding sites.** *Bioinformatics* 2005, **21**(13): 2933–2942.
80. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR: **A method and server for predicting damaging missense mutations.** *Nat Methods* 2010, **7**(4):248–249.
81. Reese MG, Eeckman FH, Kulp D, Haussler D: **Improved splice site detection in genie.** *J Comput Biol* 1997, **4**(3):311–323.
82. Goldstein DB, Ruiz Linares A, Cavalli-Sforza LL, Feldman MW: **Genetic absolute dating based on microsatellites and the origin of modern humans.** *Proc Natl Acad Sci USA* 1995, **92**(15):6723–6727.
83. Slatkin M: **A measure of population subdivision based on microsatellite allele frequencies.** *Genetics* 1995, **139**(1):457–462.
84. Farrall M, Weeks DE: **Mutational mechanisms for generating microsatellite allele-frequency distributions: an analysis of 4,558 markers.** *Am J Hum Genet* 1998, **62**(5):1260–1262.
85. Behar DM, Thomas MG, Skorecki K, Hammer MF, Bulygina E, Rosengarten D, Jones AL, Held K, Moses V, Goldstein D, Bradman N, Weale ME: **Multiple origins of ashkenazi levites: Y chromosome evidence for both near eastern and european ancestries.** *Am J Hum Genet* 2003, **73**(4):768–779.

doi:10.1186/1471-2156-14-34
Cite this article as: Bains *et al.*: Molecular diversity and population structure at the Cytochrome P450 3A5 gene in Africa. *BMC Genetics* 2013 **14**:34.